DOCUMENT RESUME

ED 167 577 TM 008 260

AUTHOR Brennan, Robert L.
TITLE Some Applications of Generalizability Theory to the Dependability of Domain-Referenced Tests. ACT Technical Bulletin No. 32.
INSTITUTION American Coll. Testing Program, Iowa City, Iowa. Research and Development Div.
PUB DATE Apr 79
NOTE 79p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, California, April, 1979) ; Not available in hard copy due to marginal legibility

EDRS PRICE MF-$0.83 Plus Postage. HC Not Available from EDRS.
DESCRIPTORS Analysis of Variance; Cost Effectiveness; *Criterion Referenced Tests; Cutting Scores; *Mastery Tests; Norm Referenced Tests; Scores; *Standard Error of Measurement; Statistical Analysis; Test Construction; *Test Interpretation; *Test Reliability
IDENTIFIERS *Classical Test Theory; *Domain Referenced Tests; Generalizability Theory; Signal Noise Ratios

ABSTRACT

Using the basic principles of generalizability theory, a psychometric model for domain-referenced interpretations is proposed, discussed, and illustrated. This approach, assuming an analysis of variance or linear model, is applicable to numerous data collection designs, including the traditional persons-crossed-with-items design, which is treated extensively here. It is shown that the appropriate error variance for domain-referenced interpretations is at least equal to and usually greater than the error variance for norm-referenced interpretations. Also, it is shown that two indices, error variance and signal/noise ratio, can be developed to reflect the dependability of a domain-referenced testing procedure. These indices are interpretable in several complementary ways, and they are easily compared with generalizability coefficients. In this paper, consideration is given not only to theoretical results but also to estimation procedures, illustrative examples, extensions to multiple facet designs, and recommendations for researchers and practitioners. Problems concerning loss functions and the setting of cutting scores are discussed. It is argued that differences between norm-referenced and domain-referenced tests are differences in score interpretation, whereas test development procedures define different domains. Venn diagrams are used to explain concepts. (Author/CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED167577

ACT TECHNICAL BULLETIN NO. 32

Some Applications of Generalizability Theory To The Dependability of Domain-Referenced Tests

by

Robert L. Brennan

American College Testing Program

Iowa City, Iowa 52243

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Robert L Brennan

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

The Research and Development Division

The American College Testing Program

P. O. Box 168, Iowa City, Iowa 52243

April, 1979

TM 008 260

SOME APPLICATIONS OF GENERALIZABILITY THEORY TO THE DEPENDABILITY OF DOMAIN-REFERENCED TESTS

by

Robert L. Brennan

American College Testing Program
Iowa City, Iowa 52243

April, 1979

A paper presented in an AERA/NCME joint symposium on "Psychometric Approaches to Domain-Referenced Testing" held in San Francisco, April, 1979.

A previous version of this paper was presented at the First Annual Johns Hopkins University, National Symposium on Educational Research, entitled "Criterion-Referenced Measurement: The State of the Art," held in Washington, D.C. on October 27, 1978.

Footnote

[1] Appreciation is extended to Dr. Michael T. Kane, who participated in most of the research reviewed herein, and who offered many insightful comments about drafts of this paper. This research was partially supported by a contract between the American College Testing Program and the Office of Naval Research, through the Navy Personnel Research and Development Center, ONR Contract No. N00123-77-C-0739.

Abstract

Using the basic principles of generalizability theory, a psychometric model for domain-referenced interpretations is proposed, discussed, and illustrated. This approach is applicable to numerous data collection designs, including the traditional persons-crossed-with-items design, which is treated extensively here.

It is shown that the appropriate error variance for domain-referenced interpretations is what Cronbach and others call $\sigma^2(\Delta)$, rather than $\sigma^2(\delta)$, which is the error variance for norm-referenced interpretations. Also, it is shown that two indices can be developed that reflect the dependability of a domain-referenced testing procedure. These indices are interpretable in several complementary ways, and they are easily compared with generalizability coefficients.

In this paper, consideration is given not only to theoretical results but also to estimation procedures, illustrative examples, extensions to multiple facet designs, and recommendations for researchers and practitioners.

"Some Applications of Generalizability Theory to the Dependability of Domain-Referenced Tests"

by

Robert L. Brennan, American College Testing Program

TABLE OF CONTENTS

		Page
I.	Introduction	1
II.	Review of Approach	4
	A. Model, Assumptions, and Variance Components for the p x i Design	5
	1. Score Effects and Mean Scores	6
	2. Defining and Interpreting Variance Components	7
	3. Estimating Variance Components	9
	B. The Error Variance $\sigma^2(\Delta)$ and Domain-Referenced Interpretations	12
	1. The Error Variance $\sigma^2(\delta)$	12
	2. The Error Variance $\sigma^2(\Delta)$	13
	3. Confidence Intervals	15
	C. Indices of Dependability for Domain-Referenced Tests	18
	1. Norm-Referenced Coefficients	18
	2. An Index of Dependability for Domain-Referenced Interpretations	20
	3. A General Purpose Index of Dependability for Domain-Referenced Interpretations	23
	4. The Indices $\Phi(\lambda)$ and Φ as Agreement Statistics	24
	5. Estimating $\Phi(\lambda)$ and Φ	27
	D. An Example with Considerations of Extensions to Other Designs	32
III.	Suggestions for Future Research	37
	A. Distributional Assumptions	37
	B. Loss Functions	38
	C. Cutting Scores	39
	D. Other Issues	40
IV.	Guidelines for Practitioners	41
	A. Nature of a Domain of Items and Parallel Tests Assumptions	41
	B. Score Interpretation and Domain Specification	43
	C. Indices of Dependability	45
	D. Variance Components	47
	E. Other Issues	48
	Appendix	49
	References	52

Some Applications of Generalizability Theory to the Dependability of Domain-Referenced Tests[1]

Introduction

Fifteen years ago, Glaser (1963) suggested that there are two types of proficiency testing, which he called norm-referenced and criterion-referenced testing. Since that time, there has been considerable effort made by many to provide a psychometric basis for using and interpreting data from criterion-referenced tests. There are indeed many approaches to this issue and related issues that have been reported in the literature; most notably, there are several approaches to issues related to reliability.

The approach to reliability issues, that is reviewed and discussed here, is based primarily upon the fundamental principles of generalizability theory, as documented by Cronbach, Gleser, Nanda, and Rajaratnam (1972), and reviewed by Brennan (1977a). Many aspects of the approach discussed here can be found in Brennan (1977b, 1978), Brennan and Kane (1977a, 1977b) and Kane and Brennan (1977). Indeed, this paper may be viewed, in part, as an integrated summary of much of the work reported in these publications. As such, this paper provides an introduction to, and review of, a specific psychometric theory for criterion-referenced testing.

In the next few paragraphs several general issues are discussed that lead to some terminological conventions adopted for the purposes of this paper. To some extent, terminology is always arbitrary; but, even so, the particular conventions adopted here are at least convenient, and hopefully appropriate, for the approach to be reviewed and discussed.

Glaser and Nitko (1971) define a criterion-referenced test as "one that is deliberately constructed to yield measurements that are directly interpretable in terms of specified performance standards" (p. 653). This is probably the best-known definition of a criterion-referenced test. However, some researchers find this terminology to be cumbersome and/or misleading. Most of their objections are concerned with the word "criterion," which is sometimes judged to be especially ambiguous. To resolve this ambiguity, some have suggested distinguishing between "content-referenced" and "criterion-referenced" testing (APA, 1974). Here, however, following Hively (1974), Millman (1974), and others, the term domain-referenced will be used rather than "criterion-referenced." The word domain is particularly appropriate here because generalizability theory is based, in part, upon principles of random sampling from a specified domain (or universe) of items and the nature of scores referenced to such a domain. The domain referred to by a domain-referenced test "may be extensive or a single, narrow objective, but it must be well defined, which means that the content and format limits must be well specified" (Millman, 1974, p. 314).

Actually, the distinction between norm-referenced and domain-referenced tests, per se, is something of a misnomer. The more appropriate distinction is between norm-referenced and domain-referenced interpretations or uses of scores resulting from a test. For example, the scores from any test might be given norm-referenced or domain-referenced interpretations. Indeed, most of the literature that treats issues of reliability for domain-referenced tests actually treats the reliability of a particular set of scores that are given a domain-referenced interpretation. In this paper, to obviate verbal complexity, we will sometimes refer to norm-referenced, domain-referenced, and mastery "testing" or "tests"; however, a more complete verbal description would refer to scores that are given (or provide) norm-referenced or domain-referenced interpretations for a particular testing procedure.

Nothing in the Glaser and Nitko (1971) definition quoted above necessitates the existence or use of a single cutting score as a "specified performance standard." However, much of the literature subsumed under the heading of criterion-referenced, or domain-referenced, measurement does, in fact, postulate the existence of a single cutting score. Since this inconsistency in terminology can lead to confusion, the term mastery test is frequently used to refer to a domain-referenced test with a single fixed cutting score.

Since Popham and Husek (1969) challenged the appropriateness of correlation coefficients as indices of reliability for domain-referenced and mastery tests, considerable effort has been devoted to developing more appropriate indices. Most of these indices have been proposed as measures of reliability; here, however, the more generic term, dependability, will be used to avoid unwarranted associations with the classical theory of reliability for norm-referenced tests.

Review of Approach

Classical test theory provides a very simple structural model of the relationship between observed, true, and error scores. However, the simplicity of the model necessitates some rather restrictive assumptions if the model is to be applied to real data. Generalizability theory liberalizes and extends classical test theory in several important respects. For example, the theory of generalizability does not necessitate the classical test theory assumption of "parallel" tests; rather, generalizability theory employs the weaker assumption of "randomly parallel" tests. Also, classical test theory assumes that errors of measurement are sampled from an undifferentiated univariate distribution. By contrast, generalizability theory allows for the existence of multiple types and sources of error through the application of analysis of variance procedures, or, more specifically, through the application of the general linear model to the dependability of measurement. Consequently, generalizability theory is applicable to a broad range of testing and evaluation studies that arise in education and psychology. The subject of this paper--namely, the dependability of domain-referenced tests--is only one such application.

Since generalizability theory is unfamiliar to many practitioners, the review provided below is intentionally didactic, to some extent. In general, the text provides an outline, without proofs, of the development of theoretical results, with emphasis placed upon the interpretation of such results. For the most part, formulas for estimating results are provided in tables; and illustrative examples are reported in sufficient detail to enable the reader to verify calculations. Also, relationships between results for norm-referenced and domain-referenced testing are treated to a limited extent. Although major emphasis is placed upon results for the simple persons-crossed-with-items design, a limited treatment of some other designs is also provided.

Model, Assumptions, and Variance Components for the p x i Design

Let us assume that the data for a particular test are derived from a random sample of n_p persons from an infinite population of persons and a random sample of n_i items from an infinite universe (or domain) of items. The observed score for any person p on any item i can be represented by the linear model:

$$X_{pi} = \mu + \mu_p^{\sim} + \mu_i^{\sim} + \mu_{pi}^{\sim} + e$$

where

μ = grand mean in the population of persons and the universe of items;

$\mu_p^{\sim}$ = score effect for person p;

$\mu_i^{\sim}$ = score effect for item i;

$\mu_{pi}^{\sim}$ = score effect for the interaction of person p and item i; and

e = experimental error.

Usually, we have only one observation for each person-item combination, which implies that the $\mu_{pi}^{\sim}$ and e effects are completely confounded. Cronbach et al. (1972) represent this confounding with the notation (pi,e). Here, in accordance with the notational conventions in Brennan (1977a), this confounding will not be indicated explicitly, and we may write this equation as:

$$X_{pi} = \mu + \mu_p^{\sim} + \mu_i^{\sim} + \mu_{pi}^{\sim} . \qquad (1)$$

The basic assumptions underlying the model in Equation 1 are the typical analysis of variance (or general linear model) assumptions which are well-documented by Cronbach et al. (1972) and by most experimental design texts. Briefly, these assumptions are that all effects in the model are assumed to be sampled independently, and the expected value of each effect over the population of persons and the universe of items is set equal to zero. Given these assumptions,

Equation 1 is the random effects model for the persons-crossed-with-items design, p x i. It is particularly important to note that this model does not require any assumptions about the distributional form for the observed scores or any score effects.

Score Effects and Mean Scores. Equation 1 provides a decomposition of the observed score for person p on item i, in terms of score effects, or components. In a similar manner, the average observed score for person p on a sample of items (i.e., a test) can be represented as:

$$X_{pI} = \mu + \mu_p^{\sim} + \mu_I^{\sim} + \mu_{pI}^{\sim}, \qquad (2)$$

where I denotes the average score for a particular sample of n_i items; and, as in Equation 1, $\mu_{pI}^{\sim}$ is completely confounded with experimental error.

Following the terminology of generalizability theory, the universe score for person p is denoted μ_p and defined as the expected value of the observed score over the universe of items. Therefore, using Equation 1

$$\mu_p = \mathcal{E}_i X_{pi} = \mu + \mu_p^{\sim}; \qquad (3)$$

or, equivalently, using Equation 2

$$\mu_p = \mathcal{E}_I X_{pI} = \mu + \mu_p^{\sim},$$

where the expectation is taken over all possible samples of size n_i from the universe of items.

Similarly, the population mean for item i is:

$$\mu_i = \mathcal{E}_p X_{pi} = \mu + \mu_i^{\sim}; \qquad (4)$$

and the population mean for a test, or set of items, I, is:

$$\mu_I = \mathcal{E}_p X_{pI} = \mu + \mu_I^{\sim}. \qquad (5)$$

Equations 4 and 5, respectively, imply that $\mu_{\underline{i}}$ is the item difficulty level for the population of persons, and $\mu_{\underline{I}}$ is the test mean in terms of proportion of items correct.

In classical test theory, the assumption of parallel tests requires that $\mu_{\underline{I}} = \mu_{\underline{J}}$, where I and J are two different sets of items, or tests. Here, however, it is not assumed that tests are classically parallel. Rather, in generalizability theory the assumption of classically parallel tests is replaced by the weaker assumption of randomly parallel tests. Two tests are randomly parallel if they both consist of random samples of the same number of items from the same universe, or domain, of items.

Defining and Interpreting Variance Components. For each of the score effects, or components, in Equation 1, there is an associated variance component. For example, the variance component for persons is denoted $\sigma^2(\underline{p})$ and defined as:

$$\sigma^2(\underline{p}) = \mathcal{E}_{\underline{p}}(\mu_{\underline{p}} - \mu)^2 = \mathcal{E}_{\underline{p}}(\mu_{\underline{p}}\sim)^2 . \qquad (6)$$

Similarly,

$$\sigma^2(\underline{i}) = \mathcal{E}_{\underline{i}}(\mu_{\underline{i}} - \mu)^2 = \mathcal{E}_{\underline{i}}(\mu_{\underline{i}}\sim)^2 , \qquad (7)$$

and

$$\sigma^2(\underline{pi}) = \mathcal{E}_{\underline{p}}\,\mathcal{E}_{\underline{i}}(X_{\underline{pi}} - \mu_{\underline{p}} - \mu_{\underline{i}} + \mu)^2 = \mathcal{E}_{\underline{p}}\,\mathcal{E}_{\underline{i}}(\mu_{\underline{pi}}\sim)^2 . \qquad (8)$$

These variance components are called random effects variance components, because they are derived from the random effects model in Equation 1.

The sum of these variance components equals the total observed score variance, denoted $\sigma^2(X_{\underline{pi}})$, and not to be confused with the expected observed score variance discussed later. Note especially that these variance components are associated with score effects for single observations. If, for example, items are scored

dichotomously, then the maximum value of any of the variance componets in Equations 6-8, is 0.25. That is, $\sigma^2(p)$, $\sigma^2(i)$, and $\sigma^2(pi)$ provide a decomposition of $\sigma^2(X_{pi})$, the variance over persons and items of set of item scores, not test scores.

The variance component, $\sigma^2(p)$, assumes a particularly important role here, in that it is the universe score variance; i.e., the variance over persons of the universe scores defined in Equation 3. In generalizability theory, the universe score variance plays a role like that of the true score variance in classical test theory. Similarly, $\sigma^2(pi)$ is analogous to the usual error variance, for a single item test, in classical test theory. Classical test theory, however, has no quantity analogous to $\sigma^2(i)$, which is by definition the variance of the mean scores for all items in the universe, where each mean is based upon the population of persons. As shown later, $\sigma^2(i)$ plays an especially important role in distinctions between norm-referenced and domain-referenced interpretations of test scores.

By substituting I for i everywhere in Equations 6-8, we obtain the definitions of the variance components associated with Equation 2. It is evident that $\sigma^2(p)$, the universe score variance, remains unchanged; and it is easily shown that

$$\sigma^2(I) = \sigma^2(i)/n_i \qquad (9)$$

and

$$\sigma^2(pI) = \sigma^2(pi)/n_i \; . \qquad (10)$$

Recall that I denotes an average score for a particular sample of n_i items. This is the usual convention in generalizability theory; but if one prefers total scores variance components, they are obtained be multiplying all variance components by n_i^2.

In Equation 10, $\sigma^2(\underline{pI})$ is the usual error variance in classical test theory, for a test of length $\underline{n}_i$. However, $\sigma^2(\underline{I})$ in Equation 9 is not analogous to any quantity in classical test theory. Indeed the assumption of parallel tests in classical test theory necessitates that $\sigma^2(\underline{I})$ be zero. To appreciate this fact, recall that the assumption of parallel tests requires that, for two different sets of items, $\underline{I}$ and $\underline{J}$, the population means be equal; i.e., $\mu_{\underline{I}} = \mu_{\underline{J}}$ (see Equation 5). When this assumption is extended to multiple sets of items (tests of length $\underline{n}_i$), it is evident that the assumption is met only if $\sigma^2(\underline{I}) = 0$. Again, generalizability theory assumes randomly parallel, not classically parallel, tests.

Estimating Variance Components. In generalizability theory, variance components assume central importance. They are the building blocks that provide a crucial foundation for all subsequent results. From a practical point of view, therefore, the estimation and interpretation of variance components is a matter of considerable importance. In understanding variance components, the defining Equations 6-8 are useful, but they cannot be used directly to estimate variance components, because these equations are expressed in terms of model parameters, not estimates.

Rather, one usually estimates variance components through an analysis of variance, such as that provided in Table 1 for the $\underline{p} \times \underline{i}$ design. It is evident from Table 1 that $\sigma^2(\underline{p})$, $\sigma^2(\underline{i})$, and $\sigma^2(\underline{pi})$ are estimated using mean squares and sample sizes, only; and once these variance components have been estimated, Equations 9 and 10 can be used to estimate $\sigma^2(\underline{I})$ and $\sigma^2(\underline{pI})$. (For present purposes, the reader should disregard all references in Table 1 to G study and D study, as well as the primes attached to some sample sizes.) Since variance components are unfamiliar statistics for many readers, the lower third of Table 1 provides equations for some relatively common statistics in terms of variance components.

Insert Table 1 about here

Variance components can also be estimated using Venn diagrams. More importantly, however, Venn diagrams provide a useful visual aid in understanding variance components, as well as other aspects of generalizability analyses. Figure 1 provides the general form of a Venn diagram for the p x i design, as well as the relationships among areas of a Venn diagram and certain equations frequently encountered in analysis of variance and/or generalizability theory. As indicated in Figures 1A, 1B, and 1C, circles and their intersections can be interpreted as mean squares, while parts of circles can be interpreted as variance components, or simple functions of them. These relationships provide a visual procedure for estimating variance components. From Figures 1B and 1D, it is evident that subtracting MS(pi) from MS(p), and dividing by n_i, provides an estimate of $\sigma^2(p)$, the universe score variance.

Insert Figure 1 about here

To further clarify the estimation and interpretation of variance components, as well as the notational conventions used here, consider the synthetic data in Table 2, which has been studied extensively by Guilford (1954) from the perspective of classical test theory. These data are introduced not because they are illustrative of any ideal characteristics of a test, but rather because these data can be studied in their entirety by the reader, without encountering either trivial results of extensive computational complexity. In particular, using the equations

in Table 1 and the data in Table 2, it is easy to verify the numerical values of the estimated variance components in Table 3. It is evident that, for these data, the estimated universe score variance, $\hat{\sigma}^2(\underline{p})$, is slightly less than estimated variance among item mean scores, $\hat{\sigma}^2(\underline{i})$; but neither of these variance components is nearly as large as $\hat{\sigma}^2(\underline{pi})$.

Insert Tables 2 and 3 about here

The Error Variance $\sigma^2(\Delta)$ and Domain-Referenced Interpretations

One of the distinct advantages of generalizability theory over classical test theory is that generalizability theory allows us to distinguish clearly between two different types of error, denoted δ and Δ. As argued below, δ and its associated variance, $\sigma^2(\delta)$, are appropriate for norm-referenced interpretations; while Δ and its variance, $\sigma^2(\Delta)$, are appropriate for domain-referenced interpretations. For the approach developed here to domain-referenced issues of the dependability of measurement, this distinction is of crucial importance. Indeed this approach suggests that classical test theory is generally incapable of distinguishing between norm-referenced and domain-referenced interpretations. We begin with a discussion of $\sigma^2(\delta)$ to enable subsequent comparisons of $\sigma^2(\Delta)$ with $\sigma^2(\delta)$.

The Error Variance $\sigma^2(\delta)$. In norm-referenced testing interest is focused on "the relative ordering of individuals with respect to their test performance, for example, whether student A can solve his problems more quickly than student B" (Glaser, 1963, p. 519), or "the adequacy of the measuring procedure for making comparative decisions" (Cronbach et al., 1972, p. 95). In this situation, the error for a given person, as defined by Cronbach et al. (1972), is:

$$\delta_p = (X_{pI} - \mu_I) - (\mu_p - \mu) \quad . \tag{11}$$

In Equation 11, μ represents the reference point for making statements about μ_p; and μ_I represents the reference point for making statements about X_{pI}. That is, the raw score, X_{pI}, carries no inherent meaning for a norm-referenced test. It is the deviation score, $X_{pI} - \mu_I$, that carries the meaning, and this deviation score is to be interpreted as an estimate of $\mu_p - \mu$.

When X_{pI}, μ_p, and μ_I are expressed in terms of score effects (see Equations 2, 3, and 5), we obtain:

$$\delta_p = \mu_{pI}^{\sim} \qquad (12)$$

with variance

$$\sigma^2(\delta) = \sigma^2(pI) . \qquad (13)$$

This is the error variance for norm-referenced interpretations of test data, and it is identical to the error variance incorporated in coefficients such as a KR-20 and Cronbach's (1951) α, although the derivation of $\sigma^2(\delta)$ is based upon generalizability theory--not classical test theory.

From Table 3, for our illustrative data, we see that our estimate of $\sigma^2(\delta)$ is $\hat{\sigma}(pI) = 0.0105$. It is well-known that the error variance in classical test theory is obtained by multiplying the observed variance by 1 - r, where r is the reliability of the test. From Table 2, for these data, the observed variance is 0.0729, KR-20 is 0.856, and, therefore, the error variance is 0.0729(1 - 0.856) = 0.0105, which is identical to $\sigma^2(\delta)$.

The Error Variance $\sigma^2(\Delta)$. In norm-referenced testing, then, our interest is in observed deviation scores, $X_{pI} - \mu_I$, which are to be interpreted relative to universe deviation scores, $\mu_p - \mu$. In domain-referenced testing, however, our interest is in "the degree to which the student has attained criterion performance" (Glaser, 1963, p. 519), independent of the performance of other students. That is, we are not primarily interested in the relative ordering of examinees' universe scores; rather, we are interested in the difference between each examinee's universe score and some absolute standard. In this case, the error for a given examinee is

$$\Delta_{pI} = (X_{pI} - \lambda) - (\mu_p - \lambda) = X_{pI} - \mu_p , \qquad (14)$$

where λ is some reference point. Here, no restrictions are placed upon the value of λ except that it be expressed as a proportion of items correct, to be consistent with the convention adopted previously of using mean scores, rather than total

scores. Usually, λ is defined by the user, a priori; and, in such cases, it is common to speak about mastery interpretations, where λ is the mastery cutting score. However, it will be shown that λ might also be expressed as a population parameter or sample statistic, for certain more general domain-referened interpretations.

When X_{pI} and μ_{p} in Equation 14 are expressed in terms of score effects (see Equations 2 and 3), we obtain

$$\Delta_{pI} = \mu_{I}\tilde{} + \mu_{pI}\tilde{} \qquad (15)$$

with variance

$$\sigma^2(\Delta) = \sigma^2(I) + \sigma^2(pI). \qquad (16)$$

This is the error variance for domain-referenced interpretations.

It is evident from Equations 13 and 16 that $\sigma^2(\Delta)$ is always at least as large as $\sigma^2(\delta)$. For example, for our illustrative data $\hat{\sigma}^2(\Delta) = 0.0061 + 0.0105 = 0.0166$, which is greater than $\hat{\sigma}^2(\delta) = 0.0105$. Indeed, the difference between $\sigma^2(\Delta)$ and $\sigma^2(\delta)$ is simply $\sigma^2(I)$. Therefore, the two error variances are equal only when μ_I is a constant for all instances of the testing procedure. The variance component for the main effect for items, $\sigma^2(I)$, reflects differences in the mean score (in the population) for different samples of items. If we are interested only in differences among examinee universe scores, as in norm-referenced testing, then any effect which is a constant for all examinees does not contribute to the error variance. However, for domain-referenced testing, we are interested in the absolute magnitude of examinee universe scores, or the magnitude compared to some externally defined cutting score. In this case, fluctuations in mean scores for samples of items do contribute to error variance.

From another point of view, since all persons take all items for the $\underline{p} \times \underline{i}$ design, the effect $\mu_{\underline{I}}$ in Equation 2 contributes to all examinees' observed scores in the same way. For domain-referenced testing, however, this does not eliminate the item effect as a source of error. For example, if we happen to select an especially easy set of items from the universe of items, our estimates of $\mu_{\underline{p}}$ and $\mu_{\underline{p}} - \lambda$ will tend to be too high. This component of error is taken into account in $\sigma^2(\Delta)$, but not in $\sigma^2(\delta)$.

Although, $\sigma^2(\Delta)$ cannot be obtained directly from classical test theory, Brennan and Kane (1977b) have shown that, for dichotomous data, $\hat{\sigma}^2(\Delta)$ is related to Lord's (1957) formula for the standard error of measurement for a person's score, which, in our notation is

$$\hat{\sigma}(\Delta|X_{\underline{pI}}) = \sqrt{\frac{X_{\underline{pI}}(1 - X_{\underline{pI}})}{n_{\underline{i}} - 1}} \quad . \tag{17}$$

In effect, $\hat{\sigma}^2(\Delta)$ is the average over persons of the square of Lord's standard error of measurement. The reader can verify this result for the data in Table 2. It is also possible to show that $\hat{\sigma}^2(\Delta)$ is an estimate of the generic error variance discussed in Lord and Novick (1968, Chapter 8).

Confidence Intervals. $\sigma^2(\Delta)$ and $\sigma^2(\delta)$ are discussed extensively by Cronbach et al. (1972), but they do not directly consider $\sigma^2(\Delta)$ in the context of domain-referenced testing. Rather, they state:

> Classical theory, assuming uniform condition means ($\mu_{\underline{I}} = \mu$), ignores the distinction between Δ and δ. Lord (1962) pointed out that condition means are unlikely to be equal when tests are not carefully equated. Lord showed that the variance of the within-person error Δ, over

> nonequivalent tests, differs from the error variance calculated by classical formulas... A confidence interval of the conventional sort has to be defined in terms of $\sigma(\Delta)$, not $\sigma(\delta)$; only if all tests (or other procedures) yield strictly equal means is it appropriate to use $\sigma(\delta)$. p. 82.

When Cronbach et al. (1972) speak of a "confidence interval of the conventional sort," they are speaking in the context of using the observed score, X_{pI}, as an estimate of the universe score, μ_p. In this context it is clear from Equation 14 that the error variance for a confidence interval is indeed $\sigma^2(\Delta)$. More generally, $\sigma^2(\Delta)$ is the appropriate error variance for a confidence interval when $X_{pI} - \lambda$ is used as an estimate of $\mu_p - \lambda$, independent of the value of λ. However, if the observed score under consideration is the deviation score, $X_{pI} - \mu_I$, which is used as an estimate of $\mu_p - \mu$, then $\sigma^2(\delta)$ is the appropriate error variance for a confidence interval. We might say, therefore, that for the score of a single examinee on a test, $\sigma^2(\Delta)$ is the appropriate error variance for confidence intervals involving domain-referenced interpretations; whereas $\sigma^2(\delta)$ is the appropriate error variance for confidence intervals involving norm-referenced interpretations.

For the illustrative data in Tables 2 and 3, a 68% confidence interval for the eighth examinee with a score of 0.75 would extend from $[0.75 - \hat{\sigma}(\Delta)]$ to $[0.75 + \hat{\sigma}(\Delta)]$. Since $\hat{\sigma}^2(\Delta)$ is 0.0166, $\hat{\sigma}(\Delta)$ equals 0.1288, and this confidence interval is 0.62 to 0.88, approximately. Alternatively, one could use Lord's standard error of measurement in Equation 17 to establish a confidence interval for any examinee with a score of 0.75. Using Equation 17, $\hat{\sigma}(\Delta|X_{pI}) = \sqrt{(0.75)(0.25)/11} = 0.1306$, and a 68% confidence interval again extends from 0.62 to 0.88, approximately. In this case, the confidence interval based upon Lord's standard error of measurement is virtually identical to interval based upon $\hat{\sigma}(\Delta)$. In general, however, the

length of such intervals will differ somewhat, because Lord's standard error of measurement varies for examinees with different scores.

It should be noted that in order to attach a probability statement to a confidence interval, it is necessary to make some distributional assumption. For such purposes it is traditional to assume a normal distribution, although binomial and compound binomial assumptions are sometimes made, especially in domain-referenced testing. As pointed out previously, such distributional assumptions are not an integral part of generalizability theory, although they are required for the use of error variances in establishing confidence intervals.

The line of reasoning discussed above for identifying the appropriate error variance for a confidence interval for a single examinee's score can be extended to a consideration of confidence intervals for: the difference between the scores for two different examinees on two different tests; the difference between the scores for the same examinee on two different tests; and the difference between the scores for two different examinees, on the same test. Each of these three cases can be considered in the context of domain-referenced interpretations and norm-referenced interpretations. Partial treatment of this issue is provided by Brennan (1978).

Indices of Dependability for Domain-Referenced Tests

Using generalizability theory it is possible to define two indices of dependability for domain-referenced tests. These indices can be given interpretations that have some similarities with interpretation of classical reliability coefficients and generalizability coefficients. However, since indices for the dependability for domain-referenced tests are indeed different from analogous indices for norm-referenced tests, we always refer to domain-referenced indices as "indices of dependability" and to norm-referenced indices as "reliability coefficients" (based on classical test theory) or "generalizability coefficients" (based on generalizability theory). A principal difference is that the domain-referenced indices discussed below incorporate $\sigma^2(\Delta)$, while the norm-referenced coefficients incorporate $\sigma^2(\delta)$. We begin with a brief discussion of norm-referenced coefficients to enable subsequent comparisons of indices of dependability with such coefficients.

Norm-Referenced Coefficients. Cronbach et al. (1972) define a generalizability coefficient for the $\underline{p} \times \underline{i}$ design as:

$$\mathcal{E}\rho^2 = \frac{\mathcal{E}_{\underline{p}}(\mu_{\underline{p}} - \mu)^2}{\mathcal{E}_{\underline{I}}\,\mathcal{E}_{\underline{p}}(X_{\underline{pI}} - \mu_{\underline{I}})^2} . \qquad (18)$$

The numerator is simply the universe score variance, $\sigma^2(\underline{p})$; and the denominator is called the expected observed score variance (see Figure 1E):

$$\mathcal{E}\sigma^2(\underline{x}) = \sigma^2(\underline{p}) + \sigma^2(\delta).$$

In terms of variance components, therefore,

$$\mathcal{E}\rho^2 = \frac{\sigma^2(\underline{p})}{\sigma^2(\underline{p}) + \sigma^2(\delta)} . \qquad (19)$$

Technically, $\mathcal{E}\rho^2$ is an intraclass correlation coefficient.

The generalizability coefficient in Equation 19 is approximately equal to the expected value of the correlation between randomly selected instances of the testing procedure, where the scores correlated are deviation scores of the form $X_{pI} - \mu_I$. (Recall that such deviation scores are inherent in the definition of δ in Equation 11.) Also, it is well known that, when the variance components in Equation 19 are replaced by their estimates, $\xi\rho^2$ is identical to Cronbach's (1951) Coefficient α, and to KR-20 when items are scored dichotomously. Therefore, this simple generalizability coefficient is interpretable in much the same way as these classical estimates of reliability, even though the derivation of $\xi\rho^2$ is based on the model Equation 2, not the classical test theory model.

The generalizability coefficient, $\xi\rho^2$, can also be expressed as a simple function of the signal/noise ratio

$$\psi(\underline{g}) = \underline{S}(\underline{g})/\underline{N}(\underline{g}) = \sigma^2(\underline{p})/\sigma^2(\delta), \qquad (20)$$

where $\underline{g}$ is used to distinguish this ratio from corresponding ratios for domain-referenced interpretations discussed later. Given Equation 20, it is easy to show that

$$\xi\rho^2 = \psi(\underline{g})/[1 + \psi(\underline{g})].$$

$$= \underline{S}(\underline{g})/[\underline{S}(\underline{g}) + \underline{N}(\underline{g})].$$

The signal/noise ratio $\psi(\underline{g})$ is itself a useful alternative coefficient for norm-referenced interpretations. The signal/noise concept arises naturally in discussing communication systems where the "signal to noise ratio compares the strength of the transmission to the strength of the interference" (Cronbach and Gleser, 1964, p. 468). The signal is intended to characterize the magnitude of

the desired discriminations. Noise characterizes the effect of extraneous variables in blurring these discriminations. If the signal is large compared to the noise, the intended discriminations are easily made. If the signal is weak compared to the noise, the intended discriminations may be completely lost. In norm-referenced testing, the signal, $\sigma^2(p)$, is a function of the magnitude of the intended discriminations, $\mu_p - \mu$. These intended discriminations reflect the tolerance, or sensitivity requirements that must be met if the measurement procedure is to achieve its intended purpose. The noise, $\sigma^2(\delta)$, reflects the degree of precision, or the magnitude of the errors that arise in the testing procedure. The signal/noise ratio is, therefore, a measure of the relative precision of the testing procedure.

An Index of Dependability for Domain-Referenced Mastery Interpretations. In our treatment of error variance for domain-referenced interpretations, we noted that the differences we want to detect are of the form $\mu_p - \lambda$, and the observed differences are of the form $X_{pI} - \lambda$, where λ is a reference point, or mastery cutting score. As shown above, for a norm-referenced test the corresponding differences, $\mu_p - \mu$ and $X_{pI} - \mu_I$, provide a basis for defining the generalizability coefficient, $\mathcal{E}\rho^2$, in Equation 19. In a similar manner, an index of dependability for domain-referenced mastery testing can be defined by substituting λ for μ and μ_I in Equation 19. The resulting index is:

$$\Phi(\lambda) = \frac{\mathcal{E}_p(\mu_p - \lambda)2}{\mathcal{E}_I \mathcal{E}_p(X_{pI} - \lambda)^2} ; \qquad (21)$$

and in terms of variance components

$$\Phi(\lambda) = \frac{\sigma^2(p) + (\mu - \lambda)^2}{\sigma^2(p) + (\mu - \lambda)^2 + \sigma^2(\Delta)} . \qquad (22)$$

Equation 22 was first derived by Brennan and Kane (1977a), who denoted this index "$\underline{M}(\underline{C})$." Since the derivation is rather involved, it will not be repeated here. We simply note that the derivation relies heavily on the fact that the numerator and denominator of Equation 21 are expected squared deviations, rather than variances. In this sense, the derivation of $\Phi(\lambda)$ in terms of variance components is similar to the derivation of Livingston's (1972) coefficient. Indeed, it can be shown that Equation 21 is identical to Livingston's coefficient when $\sigma^2(\Delta)$ equals $\sigma^2(\delta)$--i.e., when $\sigma^2(\underline{I})$ is zero. This is consistent with the fact that Livingston's coefficient is based on the assumption of classically parallel tests, which precludes distinguishing between $\sigma^2(\Delta)$ and $\sigma^2(\delta)$. By contrast, the development of $\Phi(\lambda)$ is based on the assumption of <u>randomly</u> parallel tests.

Several characteristics of $\Phi(\lambda)$ are evident from Equation 22. First and most importantly, $\Phi(\lambda)$ incorporates the error variance $\sigma^2(\Delta)$, rather than the error variance $\sigma^2(\delta)$ in $\xi\rho^2$ (Equation 19). Second, $\Phi(\lambda)$ will be different for different values of λ. Third, $\Phi(\lambda)$ has a lower limit of zero and upper limit of one. Fourth, the numerator of $\Phi(\lambda)$ is dependent upon both the universe score variance, $\sigma^2(\underline{p})$, <u>and</u> the squared distance of the population mean, μ, from the reference point or cutting score, λ. Clearly, $\Phi(\lambda)$ can be positive even when there is <u>no</u> variance among examinee universe scores. Suppose, for example, that $\sigma^2(\Delta)$ is positive and all examinees have identical "high" universe scores relative to λ. In this case, $\Phi(\lambda)$ will be positive, reflecting the fact that it is relatively easy to correctly determine whether or not examinees are above the cutting score. By contrast, in this case, $\xi\rho^2$ would be zero.

$\Phi(\lambda)$ can be interpreted in terms of signal and noise concepts in a manner analogous to one interpretation of $\xi\rho^2$ discussed previously. In particular, Brennan and Kane (1977b) show that the signal/noise ratio for domain-referenced interpretations, $\underline{d}$, is

$$\psi(\underline{d}) = \underline{S}(\underline{d})/\underline{N}(\underline{d}) = [\sigma^2(\underline{p}) + (\mu - \lambda)^2]/\sigma^2(\Delta); \quad (23)$$

and, in terms of this ratio,

$$\Phi(\lambda) = \psi(\underline{d})/[1 + \psi(\underline{d})]$$

$$= \underline{S}(\underline{d})/[\underline{S}(\underline{d}) + \underline{N}(\underline{d})].$$

Note, in particular, that for domain-referenced interpretations the noise is $\sigma^2(\Delta)$, not $\sigma^2(\delta)$; and, when such interpretations are made with respect to a mastery cutting score, λ, the signal is $\sigma^2(p) + (\mu - \lambda)^2$.

It is clear from Equation 23 that the signal/noise ratio for a domain-referenced mastery test, $\psi(\underline{d})$, can be large, even if $\sigma^2(\underline{p})$ is zero, provided $(\mu - \lambda)^2$ is large relative to $\sigma^2(\Delta)$. In other words, this signal/noise ratio will be relatively large whenever the detection of the signal, $\mu_{\underline{p}} - \lambda$, is relatively easy; or, equivalently, whenever it is easy to make correct classifications of randomly selected persons. In short, the signal/noise ratio, $\psi(\underline{d})$, is a measure of the relative precision of a domain-referenced mastery testing procedure, and the index $\Phi(\lambda)$ is a simple monotonic function of $\psi(\underline{d})$.

Since $\Phi(\lambda)$ and $\psi(\underline{d})$ are monotonically related, the choice between them (if such a choice is required) should be made on the basis of convenience or interpretability. In this regard, the reader should note that unlike $\Phi(\lambda)$, the signal/noise ratio $\psi(\underline{d})$ does not have an upper limit of one. For example, when $\Phi(\lambda)$ is 0.85, $\psi(\underline{d})$ is 0.85/0.15, or 5.67. Whether or not this characteristic of $\psi(\underline{d})$, and other signal/noise ratios, is desirable depends upon one's perspective.

A General-Purpose Index of Dependability for Domain-Referenced Interpretations. It is evident from Equation 21 that $\Phi(\lambda)$ achieves its lower limit when λ equals μ, the grand mean in the population of persons and the universe of items. That is, the lower limit of $\Phi(\lambda)$ is simply

$$\Phi = \frac{\sigma^2(p)}{\sigma^2(p) + \sigma^2(\Delta)} \quad . \tag{24}$$

In comparing Φ with $\mathcal{E}\rho^2$ we note that, since $\sigma^2(\Delta)$ is always at least as large as $\sigma^2(\delta)$, the index of dependability Φ must also be less than or equal to the generalizability coefficient $\mathcal{E}\rho^2$. Intuitively, this is a reasonable characteristic of Φ since domain-referenced interpretations of "absolute" scores are more "stringent" than norm-referenced interpretations of "relative" scores.

The index Φ can also be interpreted in terms of signal/noise concepts in a matter analogous to the interpretation of $\Phi(\lambda)$. For Φ, the noise is still $\sigma^2(\Delta)$, but the signal is $\sigma^2(p)$, resulting in the signal/noise ratio $\sigma^2(p)/\sigma^2(\Delta)$. In this case, the signal reflects universe score differences among persons, independent of any cutting score. The reasonableness of such a definition of signal for a domain-referenced test may be clarified by the following argument:

> In some cases, a person's score on a domain-referenced test may be used as a descriptive statistic. In other cases, the test may be designed to serve several types of decisions. Under either of these cases, it would be useful to have a general index of the precision of the test, in addition to any indices associated with specific decision procedures. If a particular decision procedure is not specified, the value to be assigned to the signal, or tolerance, must be somewhat vague. However, a natural and useful definition is provided by the magnitudes of the differences that do exist in the population with which the instrument is used. There is strong precedent for this choice in the traditions of physical measurement. General-purpose instruments for measuring length, for example, are typically evaluated by their ability to detect differences of the order of magnitude of those encountered in some area of practice. Thus, rulers are used in carpentry and verniers are used in machine shops (Brennan & Kane, 1977b, pp. 616-617).

The Indices $\Phi(\lambda)$ and Φ as Agreement Statistics. Previous sections have discussed various distinctions between the domain-referenced indices of dependability, $\Phi(\lambda)$ and Φ, as well as distinctions between these indices and more traditional norm-referenced coefficients. A particularly useful and informative framework for interpreting $\Phi(\lambda)$ and Φ, and for identifying similarities and differences among these indices and others, has been provided by Kane and Brennan (1977). They discuss two general agreement coefficients, one of which involves a correction for chance agreement. These two general coefficients are:

$$\theta = \frac{\underline{A}}{\underline{A}_{\underline{m}}} \tag{25}$$

and

$$\theta_{\underline{c}} = \frac{\underline{A} - \underline{A}_{\underline{c}}}{\underline{A}_{\underline{m}} - \underline{A}_{\underline{c}}}; \tag{26}$$

where $\underline{A}$ is observed agreement, $\underline{A}_{\underline{m}}$ is maximum agreement, and $\underline{A}_{\underline{c}}$ is chance agreement. Under very weak assumptions about $\underline{A}$, $\underline{A}_{\underline{m}}$, and $\underline{A}_{\underline{c}}$, it is possible to show that both coefficients have an upper limit of one.

In general, θ indicates now closely the scores for any examinee can be expected to agree on randomly sampled instances of a testing procedure. By contrast, $\theta_{\underline{c}}$ indicates how closely such scores can be expected to agree when the contribution of chance agreement is removed. To completely specify either of these coefficients, however, one must define an appropriate agreement function.

For domain-referenced interpretations, one such agreement function is:

$$\underline{d} = (X_{\underline{pI}} - \lambda)(X_{\underline{qJ}} - \lambda),$$

where p and q are person subscripts, and I and J are test subscripts. When this agreement function is used in conjunction with the mathematical definitions of A, A_m, and A_c in Kane and Brennan (1977), the results obtained are those provided in Table 4. Note in particular that the agreement coefficient $\theta(d)$ is identical to $\Phi(\lambda)$; and the agreement coefficient corrected for chance, $\theta_c(d)$, is identical to Φ.

Insert Table 4 about here

These results imply that the two indices, $\Phi(\lambda)$ and Φ, address different questions about dependability. In general, $\Phi(\lambda)$ indicates how closely, in terms of the agreement function, d, the scores for any examinee can be expected to agree. Φ indicates how closely (again, in terms of the agreement function, d) the two scores for an examinee can be expected to agree, with the contribution of chance agreement removed. The index, $\Phi(\lambda)$, therefore, characterizes the dependability of decisions, or estimates, based on the testing procedure. The magnitude of $\Phi(\lambda)$ depends, in part, on chance agreement; it may be greater than zero even when decisions based on the testing procedure are no more dependable that the decisions based on marginal probabilities in the population. The index, Φ, characterizes the contribution of the testing procedure to the dependability of the decisions, over what would be expected on the basis of chance agreement. $\Phi(\lambda)$ provides an estimate of the dependability of the decisions based on the testing procedure; Φ provides an estimate of the contribution of the testing procedure to the dependability of such decisions. The two indices provide answers to different questions. The issue is not which of these indices is best, but rather which is appropriate in a given context.

Table 5 also indicates that, the loss, defined in general as

$$\underline{L} = \underline{A}_{\underline{m}} - \underline{A},$$

is $\sigma^2(\Delta)$, for the domain-referenced agreement function $\underline{d}$. In effect, therefore, depending upon the perspective one chooses, for domain-referenced interpretations $\sigma^2(\Delta)$ can be interpreted in terms of error, noise, or loss in the measurement procedure.

The approach outlined above can also be applied to other agreement functions. For example, Table 5 shows that, for the agreement function

$$\underline{g} = (X_{\underline{pI}} - \mu_{\underline{I}})(X_{\underline{qJ}} - \mu_{\underline{J}}),$$

both θ and $\theta_{\underline{c}}$ are identical to the generalizability coefficient, $\xi\rho^2$, with the loss, $\underline{L}$, being $\sigma^2(\delta)$. In effect, $\xi\rho^2$ incorporates its own correction for chance agreement, and $\sigma^2(\delta)$ can be interpreted in terms of error, noise, or loss in a norm-referenced measurement procedure.

Both the $\underline{d}$ and $\underline{g}$ agreement functions are consistent with squared error loss approaches to measurement error. Hambleton and Novick (1973), Huynh (1976), Subkoviak (1976), and Swaminathan, Hambleton, and Algina (1974), among others, suggest using a threshold loss, or agreement, function for domain-referenced interpretations. As indicated in Table 5, for the threshold agreement function, $\underline{t}$, the coefficient θ is simply

$$\theta(\underline{t}) = \Sigma\pi_{\underline{ii}},$$

the proportion of persons consistently classified as masters or non-masters on two randomly selected instances of the testing procedure. When a correction for chance agreement is incorporated, the resulting coefficient is

$$\theta_c(t) = \frac{\Sigma\pi_{ii} - \Sigma\pi_i^2}{1 - \Sigma\pi_i^2};$$

Where $\Sigma\pi_i^2$ is the sum of the squared marginal proportions of masters and non-masters in the population. Under the assumption of randomly parallel tests $\theta_c(t)$ is identical to Cohen's (1960) coefficient kappa. From this perspective, therefore, the choice between $\Phi(\lambda)$ and Cohen's kappa as indices of dependability is primarily a choice between two different agreement functions. This issue is discussed more fully later.

Estimating $\Phi(\lambda)$ and Φ. For the most part, the estimation of the two indices of dependability for domain-referenced tests, $\Phi(\lambda)$ and Φ, involves a straightforward replacement of variance components with their estimates, as indicated in the first two column of Table 5. There is, however, one exception to this rule. An unbiased estimate of $(\mu - \lambda)^2$ is not obtained by replacing μ, the grand mean, with $\bar{X}_{PI}$, the mean over the sample of n_p persons and n_i items. Rather, Brennan and Kane (1977a) show that an unbiased estimate of $(\mu - \lambda)^2$ is $(\bar{X}_{PI} - \lambda)^2 - \hat{\sigma}^2(\bar{X}_{PI})$, where $\hat{\sigma}^2(\bar{X}_{PI})$ is the estimated variability of the sample mean.

Insert Table 5 about here

Table 5 also reports the estimate for $\mathcal{E}\rho^2$ in terms of variance components, as well as formulas in terms of sample statistics, when items are scored dichotomously. We have already pointed out that KR-20 estimates $\mathcal{E}\rho^2$ when items are scored dichotomously. By contrast, Table 5 indicates that KR-21 estimates $\Phi(\lambda)$ when λ equals $\bar{X}_{PI}$, the mean for the sample of n_p persons and n_i items. This relationship is proved in the Appendix and discussed extensively by Brennan (1977b).

The proof is quite long, but the reader can verify the equality of the two formulas for $\hat{\Phi}(\lambda = \underline{X}_{pI})$ using the data reported in Tables 2 and 3. In effect, when items are scored dichotomously, $\hat{\Phi}(\lambda)$ can be no less than KR-21. This fact provides a useful basis for considering the domain-referenced dependability of scores on tests that have been documented with classical reliability estimates, only. Similarly, when only sample statistics are available, one can use the second formula in the first row to estimate $\Phi(\lambda)$, for any cutting score, λ. For reasons discussed below, however, it is strongly suggested that whenever possible one report variance components, and estimate indices of dependability in terms of variance components.

Using the formulas in Table 5, the reader can verify the numerical results for the indices of dependability reported in Table 3, based on the illustrative data in Table 2. The results in Table 3 illustrate, for example, that

$$\hat{\Phi}(\lambda = \underline{X}_{pI}) \leq \hat{\Phi} \leq \hat{\rho}^2;$$

or, when items are scored dichotomously, as they are in this case,

$$\text{KR-21} \leq \hat{\Phi} \leq \text{KR-20}.$$

Table 3 also indicates that $\hat{\Phi}(\lambda)$ decreases as λ moves from zero to the sample mean, and increases as λ moves from the sample mean to one. This fact is displayed graphically by the dashed line in Figure 2.

Insert Figure 2 about here

Figure 2 also reports estimated values of $\Phi(\lambda)$ for $n_i' = 5$, 10, 15, and 20 items. In effect, Figure 2 graphically suggests one of the distinctions in generalizability theory between a G study and a D study. For these illustrative data one can view the G study (or "generalizability" study) in terms of the actual $n_i = 12$ items in Table 2. Some decision-maker, however, may be interested in a D study (or "decision" study) involving generalization to a different sample of n_i' items, where n_i' need not equal n_i. It is for this reason that Table 1 distinguishes between n_i and n_i'.

Figure 2 was developed using the G study and D study variance components and other results reported in Table 6. From Table 6 and Figure 2, it is evident that increasing the number of items sampled, n_i', has no effect on the universe score variance, $\sigma^2(p)$; but increasing n_i' results in decreasing $\sigma^2(I)$, $\sigma^2(pI)$, $\sigma^2(\delta)$, and $\sigma^2(\Delta)$. Furthermore, these decreases in $\sigma^2(\Delta)$ cause the indices of dependability $\Phi(\lambda)$ and Φ to increase.

Insert Table 6 about here

Another way in which G studies and D studies can differ is in terms of the data collection design. Until now, we have implicitly assumed that both the G study and the D study use the crossed design, $p \times i$. We will continue to assume that variance components have been estimated from the crossed design, but now we will assume that the D study employs a design in which each person is administered a different random sample of n_i' items. This design is denoted $i:p$, where the colon is read "nested within."

Suppose, for example, that as a result of some G study, estimated variance components are available for the $p \times i$ design; but decisions are to be made based upon administering different random samples of n_i' items to each examinee. This frequently occurs in computer-assisted test construction (Lippey, 1974) or computer generated repeatable testing (Lord, 1977). Under these circumstances, the linear model for item scores in the D study is:

$$X_{i:p} = \mu + \mu_p + \mu_{i:p} ; \qquad (27)$$

where experimental error is confounded with the item effect $\mu_{i:p}$. The model for mean scores over n_i' items is obtained by substituting I for i everywhere in Equation 27. Cronbach et al. (1972) discuss the $i:p$ D study design extensively; and Kane and Brennan (1977) extend their discussion to considerations in domain-referenced testing. Here, we simply summarize some of the more important results.

In comparing the $i:p$ design in Equation 27 with the $p \times i$ design in Equation 1, we note that the effect for items nested within persons, $\mu_{i:p}$, in the $i:p$ design represents the confounding of μ_i and μ_{pi} in the $p \times i$ design. Also, the expected value, over persons, of the observed score for the $i:p$ design equals μ, the grand mean in the *universe* of items; but for the $p \times i$ design, this expected value equals μ_I, the mean for the *sample* of n_i items.

Figure 3 provides a Venn diagram representation of the variance components and mean squares for the $i:p$ design. From Figure 3, we note that for the $i:p$ design, *both* $\sigma^2(\delta)$ and $\sigma^2(\Delta)$ equal $\sigma^2(I:p)$. Furthermore, it is easy to show that

$$\sigma^2(I:p) = \sigma^2(I) + \sigma^2(pI),$$

where variance components to the right of the equality are for the $p \times i$ design (see Equations 9 and 10). Clearly, then, given estimates of $\sigma^2(I)$ and $\sigma^2(pI)$ for a G study employing the $p \times i$ design, one can easily

estimate $\sigma^2(\delta) = \sigma^2(\Delta)$ for a D study employing the $\underline{i}:\underline{p}$ design. Furthermore, so long as items are sampled from the same universe, the universe score variance, $\sigma^2(\underline{p})$, is identical for both designs. It follows that the domain-referenced indices of dependability for the nested design are identical to those for the crossed design; and the generalizability coefficient, $\mathcal{E}\rho^2$, for the nested design is identical to Φ for the crossed design.

Insert Figure 3 about here

These results imply that the dependability of a domain-referenced testing procedure is not affected by whether the D study uses the crossed design, $\underline{p} \times \underline{i}$, or the nested design $\underline{i}:\underline{p}$. The aim of domain-referenced testing is to provide point estimates of examinee universe scores, rather than to make comparisons among examinees. The dependability of each examinee's score is determined by the number of items administered to that examinee, not by how many items or which items are administered to other examinees. Administering the same items to all examinees, in any instance of the testing procedure, improves the dependability of norm-referenced interpretations but does not improve the dependability of domain-referenced interpretations. Furthermore, the use of different samples of items for different examinees will tend to improve estimates of group means. If, therefore, domain-referenced tests are to be used for program evaluation, the selection of independent samples of items for different examinees provides <u>more</u> dependable estimates of group means without any loss in the dependability of estimates of examinees' universe scores.

An Example with Considerations of Extensions to Other Designs

In previous sections a simple set of synthetic data have been used to illustrate computational procedures and results for the $p \times i$ design, primarily, with some consideration given to the $i:p$ design. Here, the theoretical framework discussed above is applied to a large real data base for one of the instruments in the Adult Performance Level (APL) Program. The APL Program is an objectives-based assessment program, administered by the American College Testing Program (ACT), and designed to measure the life skills proficiency of adults. Rather than emphasizing purely academic knowledge, the APL Program focuses on functional competency as manifested in basic tasks that are relevant to everyday living.

The APL Program consists of six instruments, the APL Survey (ACT, 1976) and five Content Area Measures, or CAMs (ACT, 1977). All items in all instruments are scored dichotomously (correct = 1; incorrect = 0). The data base analyzed here is for the adult version of the APL Survey, only; and this data base consists of the responses of 2432 adults to each of the $n_i = 8$ items in each of the five Content Area subtests of the APL Survey. These subtests are in the areas of Community Resources, Occupational Knowledge, Consumer Economics, Health, and Government and Law. Actually, the 2432 examinees are only a subset of the entire data base; and, the analyses reported here are only illustrative of a number of analyses of these data (see Brennan, 1978). For these reasons, these analyses should not be interpreted as a validation of the APL Program.

Table 7 reports mean squares and estimated variance components for the $p \times i$ design, for each of the five content area subtests of the APL Survey. From Table 7 it is evident, for example, that there is considerably less variability among items in the Health Content Area than in the other areas. Table 8 reports variance components, error variances, summary statistics, and indices of dependability for each of these subtests, when the number of items in any subtest is $n_i' = 8$ or $n_i' = 20$.

Insert Tables 7 and 8 about here

Consider, for example, the last row of Table 8, which provides estimates of $\Phi(\lambda)$ when the cutting score is 0.75, in terms of proportion of items correct. For the APL Survey with $\underline{n}_i' = 8$ items, the values of $\hat{\Phi}(\lambda = 0.75)$ are somewhat low, by conventional standards, suggesting that domain-referenced decisions based on eight items and a cutting score of 0.75 may not be very dependable for these subtests. For this reason, ACT has developed the five Content Area Measures (CAMs) referenced above, which are considerably longer versions of each of the five subtests in the APL Survey. ACT (1977) reports, for each CAM, estimates of $\Phi(\lambda = 0.75)$ that exceed 0.90.

Also, at the present time ACT is constructing a 100-item version of the APL Survey. It is of interest, therefore, to estimate error variances and indices of dependability for the five subtests in this new version of the Survey. These estimates are also provided in Table 8 under the columns headed $\underline{n}_i' = 20$. The last row of Table 8 indicates that $\Phi(\lambda = 0.75)$ is estimated to be in the 0.80's for each content area subtest.

The discussion above has been couched in terms of indices of dependability. In this author's opinion, it is especially important to consider also the error variances, $\hat{\sigma}^2(\Delta)$, for each of these subtests, which are reported in the fifth row of Table 8. The square root of these error variances, or the standard errors, are the estimates used in establishing confidence intervals about examinees' proportion correct scores. For $\underline{n}_i' = 8$ items, these standard errors range from 0.13 to 0.16, approximately; and for $\underline{n}_i' = 20$ items, the range is about 0.08 to 0.10.

It is important to note that the above discussion does not address domain-referenced issues of dependability for the entire 40-item APL Survey. Thus far, we have considered such issues for each subtest or content area, only. By analogy with classical approaches, one might conduct a p x i analysis in which all 40 items in the Survey were not differentiated by content area. To do so, however, would confound certain estimates and induce some ambiguities, because such an approach fails to reflect the fact that items are nested within subtests in the universe. Classical test theory cannot directly take into account such nesting, but generalizability theory can. Furthermore, the approach to domain-referenced interpretations discussed above has been extended by Brennan (1978) to such designs, and others.

Specifically, for the entire 40-item APL Survey, and the data base discussed above, each person responded to every one of the $n_i = 8$ items in each of the $n_s = 5$ subtests. We denote this design p x (i:s), where, as before, the colon is to be interpreted as "nested within." Figure 4 provides the Venn diagram and structural model for this design, along with formulas for the estimated random effects G study variance components. Note that for this design there are five independently estimable sources of variance, whereas for the p x i design there are only three. A principal advantage of the p x (i:s) design is that it enables us to distinguish effects due to items within subtests from effects due to subtest means.

Insert Figure 4 about here

Table 9 reports results for four different D studies of the entire APL Survey using the $\underline{p} \times (\underline{i}:\underline{s})$ design. The reader interested in a detailed consideration of procedures for obtaining these results can refer to Brennan (1977a, 1978). Here, we simply consider some of the more important aspects of these results. Basically, Table 9 reports four different D studies--two in which the five subtests are considered random with $n_i' = 8$ and 20, and two in which the subtests are considered fixed with $n_i' = 8$ and 20. To say that these five subtests are random is to imply that we are interested in generalizing to some essentially infinite universe of potential subtests. (See Brennan 1977a for considerations in sampling from a finite universe.) By contrast, if subtests are considered fixed, then our interest is only in the actual subtests in the APL Survey. Table 9 illustrates that when subtests are considered fixed, indices of dependability are larger and error variances are smaller, than when subtests are considered random. This is not to say that subtests should be considered fixed, here. Such a decision is generally not a psychometric issue; it is a judgment that must be made by a decision-maker with a particular purpose in mind.

Insert Table 9 about here

Recall that our estimates of $\Phi(\lambda)$ for each subtest with a cutting score of 0.75 and eight items were somewhat low by conventional standards (see Table 8). From the last row of Table 9, however, we see that for the entire 40-item APL Survey, $\hat{\Phi}(\lambda = 0.75)$ is 0.89 or 0.92, depending upon whether subtests are considered random or fixed, respectively; and for a 100-item version of the Survey these

estimates are on the order of 0.95. Again, this author suggests that such results be used in conjunction with the estimates of the error variance $\sigma^2(\Delta)$. From Table 9 it is evident that these estimates are relatively small.

Suggestions for Future Research

In considering suggestions for future research, it is desirable first to summarize some of the more salient features of the approach reviewed in previous sections. Basically, this approach may be viewed as an extension of generalizability theory to domain-referenced interpretations. As such, this approach is applicable to numerous different data collection designs including the simple p x i design; and this approach enables us to make concrete and meaningful distinctions between norm-referenced and domain-referenced interpretations--distinctions that resolve a number of ambiguities in other approaches. To a large extent, such distinctions are made possible, and indeed the approach is based upon, differentiating between two different types of error variance. There are, nevertheless, a number of issues that remain to be researched. Some of these are summarized below.

Distributional Assumptions

It has been argued that one of the advantages of generalizability theory is that it is distribution-free. That is, the error variances and indices of dependability discussed above can be defined and estimated without any assumptions about the form of the distribution of universe, observed, or error scores. Nevertheless, for some purposes, such as establishing confidence intervals, distributional assumptions are required. For classical norm-referenced interpretations, it is traditional to assume that errors of measurement are normally distributed. It has been pointed out by many, however, that for domain-referenced interpretations such an assumption may not be optimal. In particular, the binomial, or compound binomial, distribution may be a better choice for domain-referenced interpretations. However, as yet, this distributional assumption has not been extensively studied in the context of the approach to domain-referenced interpretations discussed above.

Loss Functions

The approach to domain-referenced interpretations suggested here explicitly incorporates a squared error loss function. It has been noted previously that other researchers have suggested using a threshold loss function. Some issues involved in choosing a loss function for domain-referenced interpretations have been discussed by Kane and Brennan (1977). Briefly, Kane and Brennan (1977) argue that the threshold loss function is appropriate whenever the only distinction that can be made usefully is a qualitative distinction between masters and non-masters. If, however, different degrees of mastery and non-mastery exist to an appreciable extent, the threshold loss function is not appropriate because it ignores such differences.

In many (but not all) educational contexts, differences between masters and non-masters are not purely qualitative. Rather, the attribute that is measured is conceptualized as an ordinal or interval scale, and the examinees may possess the attribute to varying degrees even though a single cutting score is used to define mastery. In this context it is important that examinees who are far above or below the cutting score be classified correctly. The misclassification of such examinees is likely to cause serious losses. The misclassification of examinees whose level of ability is close to the cutting score will involve much less serious losses. Current techniques for setting the cutting score are not very precise, and the choice of a cutting score, is to some extent, arbitrary. It is, therefore, relatively less important that the testing procedure correctly classify examinees whose level of skill is close to the specified cutting score. The domain-referenced indices suggested in this report reflect these considerations.

This is not to say that squared-error loss is always preferable to threshold-loss, or vice versa. Indeed, in some contexts both seem somewhat inappropriate; and, therefore, one potential line of research involves examining the applicability

of other loss (or utility) functions. To do so, however, is by no means trivial in theory or in practice.

Cutting Scores

In previous sections no consideration has been given to how one should define a cutting score λ. Clearly, this is a topic of considerable interest (see, for example, Huynh, 1976; Meskaukas, 1976; Millman, 1973; and NCME, 1978. Some have approached this issue from the point of view of minimizing classification errors over the population. The general applicability of such procedures is questionable, in this author's opinion. Rather, in most cases, the author believes that cutting scores are best established by users of the scores and content matter specialists who are familiar with the domain of items, and the nature of decisions to be made based upon examinee performance on the test(s) under consideration. It seems somewhat contradictory to define a mastery cutting score based upon statistical criteria that minimize "something" over a population of persons when, at least supposedly, interpretation of an examinee's performance is intended to be independent of the performance of other examinees. Perhaps more importantly, however, the act of establishing cutting scores using such criteria frequently obviates the question of principal interest--namely, what constitutes acceptable performance on a domain of items for a particular decision-making purpose? This question, generally, is best answered by the decision-makers with the assistance of content matter specialists.

If one agrees with the above perspective, then there are at least two issues that need to be researched very carefully. First, we need procedures for ascertaining, from decision-makers and content matter specialists, their judgments about appropriate cutting scores. Procedures developed by Nedelsky (1954) and Angoff (1971) may be helpful in this regard; or, perhaps, more complicated procedures based upon utility theory might be examined. Second, we need procedures

for taking into account the effect of disagreement among cutting score "setters" in the dependability of domain-referenced interpretations. It is interesting to note that, in principle, generalizability theory ought to be able to address this second issue, at least to some extent. A first step in this direction has been taken by Brennan and Lockwood (1978).

Other Issues

There are a number of issues in generalizability theory that have not been studied extensively including, for example, alternative procedures for estimating variance components, especially for non-orthogonal designs; relationships between generalizability theory and other psychometric models such as latent trait theory; and issues of validity viewed from the perspective of generalizability theory. Since the approach to domain-referenced interpretations discussed above is based, in part, on generalizability theory, it follows that studying these issues may shed additional light on domain-referenced interpretations.

Also, the approach to dependability discussed above has not yet been extended to a consideration of multiple cutting scores used simultaneously in a single instance of a testing procedure. To address this issue may necessitate some multivariate considerations of considerable complexity.

Guidelines for Practitioners

If the previous section provides suggestions to those who might extend the approach suggested in this paper, then this section is addressed to those who would apply the approach as it currently exists. Of course, these two audiences overlap to some extent--and, perhaps, to a considerable extent. Indeed, practitioners frequently generate the questions that lead to research topics, and sometimes the same persons answer these questions. In any case, any set of guidelines for practitioners is likely to raise questions while it proposes suggestions. Also, the viability of such suggestions depends upon the role of the practitioners, to some extent, and certainly the context within which the practitioner exists.

Finally, in considering the applicability of a set of guidelines, it is frequently useful to know the perspective of the person making the suggestions. To some extent, this author's perspective has been communicated through the approach taken in this paper. However, the first two subsections below expand upon two issues that seem important to this author.

Nature of a Domain of Items and Parallel Tests Assumptions

In previous sections, no assumptions have been made about the nature of the domain, or universe, of items except that it is sufficiently well-defined so that one can reasonably consider the items in a test to be a sample from the domain. In particular, for the approach suggested and reviewed in this paper, there is no requirement that the item mean scores, or difficulty levels in the population, have any particular distributional form; and there is no requirement that $\mu_I - \mu_J$ for two different tests, or sets of items, I and J. In short, we have assumed randomly parallel, as opposed to classically parallel, tests.

Clearly, if all items in the universe are homogeneous in difficulty levels, then certain simplifications of results will occur. However, it is not necessary to require such conditions. Indeed, for domain-referenced testing (and many would argue for norm-referenced testing, too) the domain of items should be specified by content matter specialists, and these specialists should not be constrained to include items in the domain only if they have a constant difficulty level in some population. Now, when the items in a domain vary in difficulty, it is impossible for all tests (i.e., random samples of the same number of items) to be classically parallel. Therefore, it can be argued that the assumption of classically parallel tests is not only unnecessary, it is also frequently unreasonable, at least for domain-referenced testing.

Sometimes test developers force two or more domain-referenced tests to have equal means by discarding, revising, adding, and/or exchanging items. In this case, the items in these tests, and all similar items, may provide a reasonably good operational definition of the domain of items. However, there is a difference between a domain of items and a set of classically parallel tests consisting of items in this domain. In particular, it is not necessarily true that all possible tests consisting of an equal number of items (sampled from the domain) would be classically parallel. Only if all items in a domain are equally different can it be argued that $\mu_I = \mu$ for all tests, I. Again, the assumption of classically parallel tests seems unnecessarily restrictive and perhaps unreasonable for domain-referenced testing.

For domain-referenced testing, therefore, the assumption of randomly parallel tests seems preferable to the assumption of classically parallel tests. Some object to the assumption of randomly parallel tests in that tests are not usually

built by a strict process of random sampling from a domain of items. Even so, to paraphrase Lord and Novick (1968, p. 235), the assumption of randomly parallel tests is a useful idealization. Although the items of a particular test may not have been drawn at random, one can usually conceive of a universe of items from which the test items might have been drawn.

The discussion above has been couched in terms of random sampling from an undifferentiated domain of items. This is the simplest case to consider, but by no means the only possibility. For example, items may be randomly sampled from a stratified universe, as illustrated in the previous discussion of the p x (i:s) design. In such cases, analyses are more involved, but the argument is essentially unchanged.

Score Interpretation and Domain Specification

From the point of view of generalizability theory, a test is simply a random sample of items from some domain. Therefore, the existence of a domain (or universe of generalization) does not form a logical basis for distinguishing between norm-referenced and domain-referenced testing. Sometimes norm-referenced testing and domain-referenced testing are said to be different because different procedures are used to construct items and/or different procedures are used to identify "acceptable" items. As discussed below, such procedures are best considered as part of the process of universe definition or domain specification--not as a basis for distinguishing between different purposes for testing, or more specifically, different uses or interpretations of test scores. Indeed, in this author's opinion, the fundamental distinction between norm-referenced and domain-referenced testing is a distinction in score use and interpretation. To avoid

verbal complexity, it is common to speak of "tests" rather than "interpretations, or uses, or scores"; however, the distinction should be kept in mind.

Some writers differentiate between norm-referenced and domain-referenced tests in terms of procedures used for constructing items. Such procedures affect domain specification not score interpretation. Suppose a group of test developers uses one procedure to construct items for what they call a "domain-referenced test", and a different procedure to construct items for what they call a "norm-referenced test." Even if both tests are for the same general content area, the domain of items will almost certainly be different for the two tests. In such a case, differences in test content are attributable to differences in the two domains, or universes of generalization, not to any inherent differences in norm-referenced and domain-referenced testing, per se. Indeed, in such a case, one could provide norm-referenced and domain-referenced interpretations for scores on tests sampling either domain.

Also, a number of researchers and practitioners have suggested that differential criteria for identifying "acceptable" items provide a basis for distinguishing between norm-referenced and domain-referenced tests. For example, it is sometimes suggested that non-discriminating items are acceptable for domain-referenced tests, but unacceptable for norm-referenced tests. It is not the author's purpose here to evaluate such criteria, but it is clear that different criteria will usually produce different tests, or sets of "acceptable" items. To be more specific, once one changes the criteria for item acceptability, the domain, or universe of generalization, is likely to change. Consequently, when the criteria for item acceptability differ for norm-referenced and domain-referenced tests, the differences in the tests themselves are actually attributable to differences in the specification of the universe of generalization.

To summarize, in this author's opinion, differences between norm-referenced and domain-referenced testing are best thought of as differences in score interpretation; while different procedures for item construction and different criteria for item acceptability usually imply different domains, or universes of generalization. Once a universe of generalization is specified, and items are sampled from it to form a test, then the scores on the test can be given either a norm-referenced or a domain-referenced interpretation. It may well be that practitioners find certain universes of generalization interesting or useful for domain-referenced interpretations, primarily; while they find other universes interesting or useful for norm-referenced interpretations, primarily. Even so, it seems best to differentiate between norm-referenced and domain-referenced testing principally in terms of score interpretation. Otherwise, one is likely to confound differences in score interpretations with differences in universes of generalization, which result from different item construction and/or identification procedures. This is not to mitigate, in any way, the importance of such procedures. The point here is that such procedures affect the universe of generalization and they do not necessarily imply a distinction between norm-referenced and domain-referenced testing, per se.

Indices of Dependability

We have considered two indices of dependability, $\Phi(\lambda)$ and Φ, for domain-referenced interpretations, as well as two corresponding signal/noise ratios. As discussed previously, indices of dependability and signal/noise ratios are monotonically related by simple equations. From a technical perspective, therefore, it is of little importance whether one emphasizes indices of dependability or signal/noise ratios. However, $\Phi(\lambda)$ and Φ have characteristics that are more familiar to most practitioners; and, therefore, the following discussion will be couched in terms of indices of dependability.

Basically, three different approaches have been discussed for defining and interpreting $\Phi(\lambda)$ and Φ. These approaches are based upon developing these indices in three ways: (a) in a manner somewhat analogous to the development of a generalizability coefficient; (b) through signal/noise concepts; and (c) as special cases of two general agreement coefficients. In terms of interpretability, this author suggests that the "agreement" approach may be the easiest for most practitioners. However, interpretations in terms of signal/noise concepts may be more informative for persons familiar with measurement procedures in the physical sciences. Finally, persons familiar with generalizability theory may find the first approach to be the most straightforward.

No matter which approach practitioners prefer in interpreting $\Phi(\lambda)$ and Φ, it is crucial that practitioners differentiate between $\Phi(\lambda)$ and Φ. These two indices address <u>different</u> questions about the dependability of decisions based on the testing procedure; Φ provides an estimate of the <u>contribution</u> of the testing procedure to the dependability of such decisions. Therefore, both indices should be reported, and practitioners should be careful not to confuse them.

Actually, of course, there are an infinite number of possible values for $\Phi(\lambda)$, depending upon the cutting score, λ. For many domain-referenced tests, therefore, it is best to report as many values of $\Phi(\lambda)$ as there are cutting scores that might be used in particular contexts. Better still, one might report a $\Phi(\lambda)$ "curve" (see, for example, Figure 2). Indeed, several curves might be reported if the number of items, in any given instance of the testing procedure, can be altered.

In interpreting the magnitudes of any specific values of $\Phi(\lambda)$ and Φ practitioners should keep in mind that $\Phi(\lambda)$ is dependent upon the universe score variance, $\sigma^2(\underline{p})$, the error variance, $\sigma^2(\Delta)$, and the quantity, $(\mu - \lambda)^2$; and Φ is dependent upon $\sigma^2(\underline{p})$ and $\sigma^2(\Delta)$. Interpretations based solely upon the magnitude of $\Phi(\lambda)$ or Φ seldom provide a sufficient basis for good decision-making.

Variance Components

In previous sections, considerably more space has been devoted to indices of dependability than to variance components. Such indices have practical as well as theoretical value. However, variance components and associated error variances are likely to be the most important results, because they frequently provide the most important basis for making decisions. Therefore, the central role that variance components play, in domain-referenced interpretations, should not be overlooked. Indeed, according to the "Standard for Educational and Psychological Tests,"

> (t)he estimation of clearly labeled components of score variance is the most informative outcome of a reliability study, both for the test developer wishing to improve the reliability of his instrument and for the user desiring to interpret test scores with maximum understanding (p. 49).

For these reasons, it is recommended that variance components, and especially the error variances, $\sigma^2(\Delta)$, be reported in documenting the technical characteristics of a domain-referenced test. Furthermore, it is recommended that users give at least as much consideration to variance components as they give to indices of dependability in making judgments about the applicability of domain-referenced tests in specific contexts, for specific purposes. Finally, it is recommended that test developers carefully consider variance components, and

especially error variance, in their deliberations about the optimal length of a domain-referenced test. Minimizing error variance is by no means the only relevant consideration in defining test length, but it is one important consideration.

Other Issues

Decisions about the appropriateness of particular loss functions cannot be made on the basis of a test, per se. Such decisions are intimately involved with a consideration of the purpose and context of the testing procedure. For this reason, it is impossible to make a statement about the universal applicability of either squared error loss functions, or threshold loss functions, for domain-referenced interpretations. Prudence would suggest, therefore, that when ambiguity exists with respect to choice of a loss function, practitioners should interpret results cautiously. Indeed, it is sometimes advisable to report and interpret results based upon more than one loss function, when ambiguity in choice of loss function exists.

In previous sections, an effort has been made to demonstrate relationships between sample statistics and other well known quantities, on one hand, and $\Phi(\lambda)$, Φ, and $\sigma^2(\Delta)$, on the other. For example, Table 5 illustrates that $\hat{\Phi}(\lambda = \bar{X}_{pI})$ equals KR-21. Such relationships are not of fundamental theoretical importance, but they can be used advantageously as a "rough and ready" way to assess the dependability of some domain-referenced interpretations. Indeed this is the principal reason why such results have been reported here. Again, however, it is strongly recommended that whenever possible one report variance components, and estimate indices of dependability, in terms of variance components.

Appendix

Proof that KR-21 Equals $\hat{\Phi}(\lambda = \underline{X}_{\underline{pI}})$

The usual computing formula for KR-21 is

$$\text{KR-21} = \left(\frac{\underline{n}_i}{\underline{n}_i - 1} \right) \left[1 - \frac{\underline{X}_{\underline{pI}}(1 - \underline{X}_{\underline{pI}})}{\underline{n}_i \, \underline{S}^2(\underline{X}_{\underline{pI}})} \right] .$$

Now,

$$\underline{X}_{\underline{pI}}(1 - \underline{X}_{\underline{pI}}) = \frac{1}{\underline{n}_p \underline{n}_i} \left[\sum_{\underline{p}} \sum_{\underline{i}} (\underline{X}_{\underline{pi}} - \underline{X}_{\underline{pI}})^2 \right] = \frac{\underline{SS}\ (\underline{total})}{\underline{n}_p \underline{n}_i} ,$$

and $$\underline{S}^2(\underline{X}_{\underline{pI}}) = \frac{1}{\underline{n}_p \underline{n}_i} \left[\underline{n}_i \sum_{\underline{p}} (\underline{X}_{\underline{pI}} - \underline{X}_{\underline{pI}})^2 \right] = \frac{\underline{SS}\ (\underline{p})}{\underline{n}_p \underline{n}_i} ,$$

where $\underline{SS}$ means "sum of squares." Therefore,

$$\text{KR-21} = \left(\frac{\underline{n}_i}{\underline{n}_i - 1} \right) \left[1 - \frac{\underline{SS}\ (\underline{total})}{\underline{n}_i \, \underline{SS}(\underline{p})} \right]$$

$$= \frac{\underline{n}_i \, \underline{SS}(\underline{p}) - \underline{SS}\ (\underline{total})}{(\underline{n}_i - 1) \, \underline{SS}(\underline{p})}$$

$$= \frac{(\underline{n}_i - 1) \, \underline{SS}(\underline{p}) - \underline{SS}(\underline{i}) - \underline{SS}(\underline{pi})}{(\underline{n}_i - 1) \, \underline{SS}(\underline{p})} ;$$

and dividing the numerator and denominator by $(n_i - 1)(n_p - 1)$ we obtain

$$KR\text{-}21 = \frac{MS(p) - MS(pi) - MS(i)/(n_p - 1)}{MS(p)}, \quad \text{(A1)}$$

where MS means "mean square."

Now, from the usual formulas for expected mean squares in a persons-crossed-with-items design (see Figure 1),

$$MS(p) = \hat{\sigma}^2(pi) + n_i\hat{\sigma}^2(p), \quad \text{(A2)}$$

$$MS(i) = \hat{\sigma}^2(pi) + n_p\hat{\sigma}^2(i), \quad \text{(A3)}$$

and

$$MS(pi) = \hat{\sigma}^2(pi). \quad \text{(A4)}$$

When Equations A2, A3, and A4 are substituted in Equation A1,

$$KR\text{-}21 = \frac{n_i\hat{\sigma}^2(p) - [n_p\hat{\sigma}^2(i) + \hat{\sigma}^2(pi)]/(n_p - 1)}{n_i\hat{\sigma}^2(p) + \hat{\sigma}^2(pi)}$$

$$= \frac{\hat{\sigma}^2(p) - \dfrac{n_p\hat{\sigma}^2(i) + \hat{\sigma}^2(pi)}{n_i(n_p - 1)}}{\hat{\sigma}^2(p) + \dfrac{\hat{\sigma}^2(pi)}{n_i}}. \quad \text{(A5)}$$

When we multiply the numerator and denominator of Equation A5 by $(n_p - 1)/n_p$, and add and subtract $\hat{\sigma}^2(\Delta) = \hat{\sigma}^2(i)/n_i + \hat{\sigma}^2(pi)/n_i$ in the denominator, we obtain

$$KR\text{-}21 = \frac{\hat{\sigma}^2(p) - \frac{\hat{\sigma}^2(p)}{n_p} - \frac{\hat{\sigma}^2(i)}{n_i} - \frac{\hat{\sigma}^2(pi)}{n_p n_i}}{\left[\hat{\sigma}^2(p) - \frac{\hat{\sigma}^2(p)}{n_p} - \frac{\hat{\sigma}^2(i)}{n_i} - \frac{\hat{\sigma}^2(pi)}{n_p n_i}\right] + \left[\hat{\sigma}^2(\Delta)\right]} . \qquad (A6)$$

We can simplify Equation A6 by noting that

$$\sigma^2(X_{pI}) = \mathcal{E}(X_{pI} - \mu)^2 = \frac{\sigma^2(p)}{n_p} + \frac{\sigma^2(i)}{n_i} + \frac{\sigma^2(pi)}{n_p n_i} , \qquad (A7)$$

where the expectation is taken over both samples of n_p persons and n_i items. When Equation A7 is expressed in terms of estimated values and replaced in Equation A6, we obtain

$$KR\text{-}21 = \frac{\hat{\sigma}^2(p) - \hat{\sigma}^2(X_{pI})}{[\hat{\sigma}^2(p) - \hat{\sigma}^2(X_{pI})] + \hat{\sigma}^2(\Delta)} ,$$

which is identical to $\hat{\Phi}(\lambda = X_{pI})$.

References

American College Testing Program. User's guide: Adult APL survey. Iowa City, Iowa: Author, 1976.

American College Testing Program. Technical supplement: APL content area measures. Iowa City, Iowa: Author, 1977.

American Psychological Association. Standards for educational & psychological tests (rev. ed.). Washington, D. C.: American Psychological Association, 1974.

Angoff, W. H. Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.). Educational Measurement. Washington, D.C.: American Council on Education, 1971.

Brennan, R. L. Generalizability analyses: Principles and procedures. ACT Technical Bulletin No. 26. Iowa City, Iowa: The American College Testing Program. Also, NPRDC Technical Report. San Diego, CA: Navy Personnel Research and Development Center, September 1977. (a)

Brennan, R. L. KR-21 and lower limits of an index of dependability for mastery tests. ACT Technical Bulletin No. 27. Iowa City, Iowa: American College Testing Program, December 1977. (b)

Brennan, R. L. Extensions of generalizability theory to domain-referenced testing. ACT Technical Bulletin No. 30. Iowa City, Iowa: American College Testing Program. Also, NPRDC Technical Report. San Diego, CA: Navy Personnel Research and Development Center, June 1978.

Brennan, R. L., & Kane, M. T. An index of dependability for mastery tests. Journal of Educational Measurement, 1977, 14, 277-289. (a)

Brennan, R. L., & Kane, M. T. Signal/noise ratios for domain-referenced tests. Psychometrika, 1977, 42, 609-625. (b)

Brennan, R. L., & Lockwood, R. E. Sources of disagreement in two cutting score procedures. A paper presented at the 1979 Annual Meeting of the National Council on Measurement in Education held in San Francisco, April 1979.

Cohen, J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 1960, 20, 37-46.

Cronbach, L. J. Coefficient alpha and the internal structure of tests. Psychometrika, 1951, 16, 292-334.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley, 1972.

Glaser, R. Instructional technology and the measurement of learning outcomes: Some questions. American Psychologist, 1963, 18, 519-521.

Glaser, R., & Nitko, A. J. Measurement in learning and instruction. In R. L. Thorndike (Ed.)., Educational Measurement (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Guilford, J. P. Psychometric methods (2nd ed.). New York: McGraw Hill, 1954.

Hambleton, R. K., & Novick, M. R. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 1973, 10, 159-170.

Hively, W. Introduction to domain-referenced testing. In W. Hively (Ed.). Domain-referenced testing. Englewood Cliffs, N.J.: Educational Technology Publications, 1974.

Huynh, H. On consistency of decisions in criterion-referenced testing. Journal of Educational Measurement, 1976, 13, 265-275.

Kane, M. T., & Brennan, R. L. Agreement coefficients as indices of dependability for domain-referenced tests. ACT Technical Bulletin No. 28. Iowa City, Iowa: American College Testing Program. Also, NPRDC Technical Report. San Diego, CA: Navy Personnel Research and Development Center, December 1977.

Lippey, G. (Ed.). Computer-assisted test construction. Englewood Cliffs, N.J.: Educational Technology Publications, 1974.

Livingston, S. A. Criterion-referenced applications of classical test theory. Journal of Educational Measurement, 1972, 9, 13-26.

Lord, F. M. Do tests of the same length have the same standard error of measurement? Educational and Psychological Measurement, 1957, 17, 510-521.

Lord, F. M. Some item analysis and test theory for a system of computer-assisted test construction for individualized instruction. Applied Psychological Measurement, 1977, 1, 511-512.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Meskauskas, J. A. Evaluation models for criterion-referenced testing: Views regarding mastery and standard-setting. Review of Education Research, 1976, 46, 133-158.

Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.). Evaluation in Education. Berkeley, CA: McCutchan, 1974.

Millman, J. Passing scores and test lengths for domain-referenced tests. Review of Educational Research, 1973, 43, 205-216.

National Council on Measurement in Education. Special issue on standard-setting. Journal of Educational Measurement, 1978, 15 (4).

Nedelsky, L. Absolute grading standards for objective tests. Educational and Psychological Measurement, 1954, 14, 3-19.

Popham, W. J., & Husek, T. R. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 6, 1-9.

Subkoviak, M. J. Estimating reliability from a single administration of a mastery test. Journal of Educational Measurement, 1976, 13, 265-276.

Swaminathan, H., Hambleton, R. K., & Algina, J. Reliability of criterion-referenced tests: A decision theoretic formulation. Journal of Educational Measurement, 1974, 11, 263-267.

Table 1

ANOVA for Estimating Variance Components

for the Random Effects $p \times i$ Design

Effect, Component, or Source	df	SS	MS	Estimated G Study Variance Component	D Study Sampling Frequency	Estimated D Study Variance Component
Persons (p)	$n_p - 1$	$SS(p)$	$MS(p)$	$\hat{\sigma}^2(p)$	1	$\hat{\sigma}^2(p)$
Items (i)	$n_i - 1$	$SS(i)$	$MS(i)$	$\hat{\sigma}^2(i)$	n_i'	$\hat{\sigma}^2(I)$
Interaction (pi)	$(n_p-1)(n_i-1)$	$SS(pi)$	$MS(pi)$	$\hat{\sigma}^2(pi)$	n_i'	$\hat{\sigma}^2(pI)$

$$\hat{\sigma}^2(p) = [MS(p) - MS(pi)]/n_i$$

$$\hat{\sigma}^2(i) = [MS(i) - MS(pi)]/n_p$$

$$\hat{\sigma}^2(pi) = MS(pi)$$

$$\hat{\sigma}^2(I) = \hat{\sigma}^2(i)/n_i$$

$$\hat{\sigma}^2(pI) = \hat{\sigma}^2(pi)/n_i$$

Variance of Observed Person Mean Scores: $\hat{S}^2_p(X_{pI}) = \hat{\sigma}^2(p) + \hat{\sigma}^2(pi)/n_i$

Variance of Observed Item Difficulty Levels: $\hat{S}^2_i(X_{pi}) = \hat{\sigma}^2(i) + \hat{\sigma}^2(pi)/n_p$

Sum of Observed Item Variances: $\sum_i \hat{S}^2_p(X_{pi}) = \hat{\sigma}^2(p) + \hat{\sigma}^2(pi) = \sum_i X_{pi}(1 - X_{pi})$ for dichotomous items

Note. Upper case P and I represent mean scores, not total scores.

Table 2

Synthetic Data Reported by Guilford (1954, p. 381)

Person	1	2	3	4	5	6	7	8	9	10	11	12	Total	X_{pI}
1	1	0	1	0	0	0	0	0	0	0	0	0	2	0.1667
2	1	1	1	0	0	1	0	0	0	0	0	0	4	0.3333
3	1	1	1	1	0	0	0	0	0	0	0	0	4	0.3333
4	1	1	0	1	1	0	0	1	0	0	0	0	5	0.4167
5	1	1	1	1	1	0	0	0	0	0	0	0	5	0.4167
6	1	1	1	0	1	1	1	0	0	0	0	0	6	0.5000
7	1	1	1	1	1	1	1	0	0	0	0	0	7	0.5833
8	1	1	1	1	0	1	1	1	1	1	0	0	9	0.7500
9	1	1	1	1	1	1	1	1	1	1	1	0	11	0.9167
10	1	1	1	1	1	1	1	1	1	1	1	1	12	1.0000
Total	10	9	9	7	6	6	5	4	3	3	2	1		
X_{Pi}	1.0	0.9	0.9	0.7	0.6	0.6	0.5	0.4	0.3	0.3	0.2	0.1	X_{PI} =	0.5417

$X_{PI}(1 - X_{PI}) = 0.2483$ $\quad \sum_i X_{Pi}(1 - X_{Pi}) = 2.03$ $\quad \hat{S}^2_i(X_{Pi}) = 0.0863$ $\quad \hat{S}^2_p(X_{pI}) = 0.0729$

Coefficient α = KR-20 = 0.856 $\quad$ KR-21 = 0.747 $\quad S^2_p(X_{pI}) = 0.0656$

Note. Upper case P and I represent mean scores, not total scores.

Table 3

G Study and D Study for Random Effects $\underline{p} \times \underline{i}$ Design Using Guilford Data and Assuming $n_{\underline{i}} = n'_{\underline{i}} = 12$

Effect, Component, or Source	df	SS	MS	Estimated G Study Variance Component	D Study Sampling Frequency	Estimated D Study Variance Component
Persons ($\underline{p}$)	9	7.875	0.8750	$\hat{\sigma}^2(\underline{p}) = 0.0625$	1	$\hat{\sigma}^2(\underline{p}) = 0.0625$
Items ($\underline{i}$)	11	9.492	0.8629	$\hat{\sigma}^2(\underline{i}) = 0.0737$	12	$\hat{\sigma}^2(\underline{I}) = 0.0061$
Interaction ($\underline{pi}$)	99	12.425	0.1255	$\hat{\sigma}^2(\underline{pi}) = 0.1255$	12	$\hat{\sigma}^2(\underline{pI}) = 0.0105$

$\hat{\sigma}^2(\delta) = 0.0105$	$\hat{\rho}^2 = 0.856$ (5.95)	$\hat{\Phi}(\lambda = 0.4) = 0.807$(4.17)	$\hat{\Phi}(\lambda = 0.7) = 0.817$(4.47)
$\hat{\sigma}^2(\Delta) = 0.0166$	$\hat{\Phi} = 0.790$ (3.77)	$\hat{\Phi}(\lambda = 0.5) = 0.754$(3.06)	$\hat{\Phi}(\lambda = 0.8) = 0.875$(6.98)
$\underline{X}_{\underline{pI}} = 0.5417$		$\hat{\Phi}(\lambda = 0.6) = 0.760$(3.16)	$\hat{\Phi}(\lambda = 0.9) = 0.915$(10.69)
$\hat{\sigma}^2(\underline{X}_{\underline{pI}}) = 0.0134$		$\hat{\Phi}(\lambda = \underline{X}_{\underline{pI}} = 0.747$(2.96)	

Note. Numbers in parentheses following values for $\hat{\Phi}(\lambda)$ are the estimated signal/noise ratios. All estimates of $\sigma^2(\underline{X}_{\underline{pI}})$ and $\Phi(\lambda)$ assume $n_{\underline{p}} = n'_{\underline{p}} = 10$.

Table 4

Coefficients for Different Agreement Functions

Agreement Function	Parameters	Agreement Coefficients
Domain-Referenced:		
$\underline{d} = (X_{\underline{pI}} - \lambda)(X_{\underline{qJ}} - \lambda)$	$\underline{A}(\underline{d}) = (\mu - \lambda)^2 + \sigma^2(\underline{p})$ $\underline{A}_{\underline{m}}(\underline{d}) = (\mu - \lambda)^2 + \sigma^2(\underline{p}) + \sigma^2(\Delta)$ $\underline{A}_{\underline{c}}(\underline{d}) = (\mu - \lambda)^2$ $\underline{L}(\underline{d}) = \sigma^2(\Delta)$	$\theta(\underline{d}) = \dfrac{(\mu - \lambda)^2 + \sigma^2(\underline{p})}{(\mu - \lambda)^2 + \sigma^2(\underline{p}) + \sigma^2(\Delta)}$ $\theta_{\underline{c}}(\underline{d}) = \dfrac{\sigma^2(\underline{p})}{\sigma^2(\underline{p}) + \sigma^2(\Delta)}$
Norm-Referenced:		
$\underline{g} = (X_{\underline{pI}} - \mu_{\underline{I}})(X_{\underline{qJ}} - \mu_{\underline{J}})$	$\underline{A}(\underline{g}) = \sigma^2(\underline{p})$ $\underline{A}_{\underline{m}}(\underline{g}) = \sigma^2(\underline{p}) + \sigma^2(\delta)$ $\underline{A}_{\underline{c}}(\underline{g}) = 0$ $\underline{L}(\underline{g}) = \sigma^2(\delta)$	$\theta(\underline{g}) = \theta_{\underline{c}}(\underline{g}) = \dfrac{\sigma^2(\underline{p})}{\sigma^2(\underline{p}) + \sigma^2(\delta)}$
Threshold:		
$\underline{t} = \begin{cases} 1 \text{ if } X_{\underline{pI}} \text{ and } X_{\underline{qJ}} \text{ result in same classification; and} \\ 0 \text{ otherwise} \end{cases}$	$\underline{A}(\underline{t}) = \Sigma\pi_{\underline{ii}}$ $\underline{A}_{\underline{m}}(\underline{t}) = 1$ $\underline{A}_{\underline{c}}(\underline{t}) = \Sigma\pi_{\underline{i}}^2$ $\underline{L}(\underline{t}) = \Sigma\pi_{\underline{ij}}$ $(\underline{i} \neq \underline{j})$	$\theta(\underline{t}) = \Sigma\pi_{\underline{ii}}$ $\theta_{\underline{c}}(\underline{t}) = \dfrac{\Sigma\pi_{\underline{ii}} - \Sigma\pi_{\underline{i}}^2}{1 - \Sigma\pi_{\underline{i}}^2}$

Note. Here p and q are different persons; I and J are different random samples of items.

Table 5

Estimates of Indices of Dependability

Estimate	In Terms of Estimated Variance Components	In Terms of Sample Statistics for a Full-Length Test of Dichotomously Scored Items
$\hat{\Phi}(\lambda)$	$= \dfrac{\hat{\sigma}^2(p) + (\bar{X}_{pI} - \lambda)^2 - \hat{\sigma}^2(\bar{X}_{pI})}{\hat{\sigma}^2(p) + (\bar{X}_{pI} - \lambda)^2 - \hat{\sigma}^2(\bar{X}_{pI}) + \hat{\sigma}^2(\Delta)}$	$= 1 - \dfrac{1}{n_i - 1}\left[\dfrac{\bar{X}_{pI}(1 - \bar{X}_{pI}) - s^2(\bar{X}_{pI})}{(\bar{X}_{pI} - \lambda)^2 + s^2(\bar{X}_{pI})}\right]$
$\hat{\Phi}(\lambda = \bar{X}_{pI})$	$= \dfrac{\hat{\sigma}^2(p) - \hat{\sigma}^2(\bar{X}_{pI})}{\hat{\sigma}^2(p) - \hat{\sigma}^2(\bar{X}_{pI}) + \hat{\sigma}^2(\Delta)}$	$= \text{KR-21} = 1 - \dfrac{1}{n_i - 1}\left[\dfrac{\bar{X}_{pI}(1 - \bar{X}_{pI}) - s^2(\bar{X}_{pI})}{s^2(\bar{X}_{pI})}\right]$
$\hat{\Phi}$	$= \dfrac{\hat{\sigma}^2(p)}{\hat{\sigma}^2(p) + \hat{\sigma}^2(\Delta)}$	
$\widehat{\xi\rho^2}$	$= \dfrac{\hat{\sigma}^2(p)}{\hat{\sigma}^2(p) + \hat{\sigma}^2(\delta)}$	$= \text{KR-20} = 1 - \dfrac{1}{n_i - 1}\left[\dfrac{\sum_i (\bar{X}_{pi})(1 - \bar{X}_{pi})/n_i - s^2(\bar{X}_{pI})}{s^2(\bar{X}_{pI})}\right]$

Note. $\hat{\sigma}^2(\bar{X}_{pI}) = \hat{\sigma}^2(p)/n'_p + \hat{\sigma}^2(i)/n'_i + \hat{\sigma}^2(pi)/n'_p n'_i$

58

Table 6

D Studies for Random Effects $\underline{p} \times \underline{i}$ Design Using Guilford Data for $n_{\underline{i}}' = 5, 10, 15, 20$

Effect, Component, or Source	Estimated G Study Variance Component	D Study Variance Components			
		$n_{\underline{i}}' = 5$	$n_{\underline{i}}' = 10$	$n_{\underline{i}}' = 15$	$n_{\underline{i}}' = 20$
Persons ($\underline{p}$)	0.0625	0.0625	0.0625	0.0625	0.0625
Items ($\underline{i}$)	0.0737	0.0147	0.0074	0.0049	0.0037
Interaction ($\underline{pi}$)	0.1255	0.0251	0.0126	0.0084	0.0063
$X_{\underline{pI}} = 0.5417$	$\hat{\sigma}^2(\delta)$ =	0.0251	0.0126	0.0084	0.0063
	$\hat{\sigma}^2(\Delta)$ =	0.0398	0.0200	0.0133	0.0100
	$\hat{\sigma}^2(X_{\underline{pI}})$ =	0.0235	0.0149	0.0120	0.0106
	$\widehat{\mathcal{E}\rho^2}$ =	0.714	0.832	0.882	0.908
	$\hat{\Phi}$ =	0.611	0.758	0.825	0.862
	$\hat{\Phi}(\lambda = X_{\underline{pI}})$ =	0.495	0.704	0.792	0.838
	$\hat{\Phi}(\lambda = 0.7)$ =	0.617	0.784	0.850	0.885
	$\hat{\Phi}(\lambda = 0.8)$ =	0.727	0.851	0.898	0.922
	$\hat{\Phi}(\lambda = 0.9)$ =	0.808	0.898	0.931	0.948

Note. All estimates of $\sigma^2(X_{\underline{pI}})$ and $\Phi(\lambda)$ assume $n_{\underline{p}} = n_{\underline{p}}' = 10$.

Table 7

G Studies for the Five Subtests of the APL Survey

Using the p x i Design[a]

Effect	df	Mean Squares					Estimated G Study Variance Components				
		s_1 = CR	s_2 = OK	s_3 = CE	s_4 = H	s_5 = GL	s_1 = GL	s_2 = OK	s_3 = CE	s_4 = H	s_5 = GL
p	2431	.38121	.52279	.50769	.65299	.54843	.0322	.0459	.0440	.0631	.0477
i	7	50.74109	62.34850	79.60269	18.03560	84.19510	.0208	.0256	.0327	.0074	.0346
pi	17017	.12371	.15540	.15548	.14858	.16689	.1237	.1554	.1555	.1486	.1669

[a] The five subtests are as follows: s_1 = Community Resources (CR); s_2 = Occupational Knowledge (OK); s_3 = Consumer Economics (CE); s_4 = Health (H); and s_5 = Government and Law (GL).

Table 8

D Studies for the Five APL Content Areas Using the p x i Design with $n_i = 8, 20$

	D Study Variance Components									
	Community Resources (CR)		Occupational Knowledge (OK)		Consumer Economics (CE)		Health (H)		Government and Law (GL)	
Effect	$n_i' = 8$	$n_i' = 20$	$n_i' = 8$	$n_i' = 20$	$n_i' = 8$	$n_i' = 20$	$n_i' = 8$	$n_i' = 20$	$n_i' = 8$	$n_i' = 20$
p	.0322	.0322	.0459	.0459	.0440	.0440	.0631	.0621	.0477	.0477
i	.0026	.0010	.0032	.0013	.0041	.0016	.0009	.0004	.0043	.0017
pi	.0155	.0062	.0194	.0078	.0194	.0078	.0186	.0074	.0209	.0084
$\hat{\sigma}^2(\delta)$ =	.0155	.0062	.0194	.0078	.0194	.0078	.0186	.0074	.0209	.0084
$\hat{\sigma}^2(\Delta)$ =	.0181	.0072	.0226	.0091	.0235	.0094	.0195	.0078	.0252	.0101
$\hat{E}\rho^2$ =	.675	.839	.703	.855	.694	.849	.772	.895	.695	.850
$\hat{\Phi}$ =	.640	.817	.670	.835	.652	.824	.764	.890	.654	.825
$\underline{X}_{\underline{pI}}$ =	.7753	.7753	.6615	.6615	.6472	.6472	.6780	.6780	.5694	.5694
$\hat{\sigma}^2(\underline{X}_{\underline{pI}})$ =	.0026	.0010	.0032	.0013	.0041	.0016	.0009	.0004	.0043	.0017
$\hat{\Phi}(\lambda = \underline{X}_{\underline{pI}})$ =	.620	.812	.654	.831	.629	.819	.761	.889	.633	.820
$\hat{\Phi}(\lambda = .50)$ =	.853	.937	.753	.886	.724	.872	.824	.924	.657	.834
$\hat{\Phi}(\lambda = .75)$ =	.625	.816	.691	.852	.682	.849	.776	.897	.751	.886

Note. All estimates of $\sigma^2(\underline{X}_{\underline{pI}})$ and $\Phi(\lambda)$ assume $n_{\underline{p}} = n'_{\underline{p}} = 2432$.

Table 9

D Studies for APL Survey Using Design p x (i:s)

Effect[a]	MS	G Study Variance Component		D Study Variance Components: Subtests Random ($n_s' = 5$)		Subtests Fixed ($n_s' = 5$)	
				$n_i' = 8$	$n_i' = 20$	$n_i' = 8$	$n_i' = 20$
p (2432)	1.84829	.0414		.0414	.0414	.0414	.0414
s (5)	106.02210	.0024		.0005	.0005		
i:s (8)	58.98439	.0242		.0006	.0002	.0006	.0002
ps	.19105	.0053		.0011	.0011	.0011	.0011
pi:s	.14857	.1486		.0037	.0015	.0037	.0015
		$\hat{\sigma}^2(\delta)$	=	.0048	.0026	.0037	.0015
		$\hat{\sigma}^2(\Delta)$	=	.0059	.0033	.0043	.0017
		$\hat{\sigma}^2(\tau)$[b]	=	.0414	.0414	.0425	.0425
		$\hat{E\rho}^2$	=	.896	.941	.920	.966
		$\hat{\Phi}$	=	.875	.926	.908	.962
		$X_{PI:S}$	=	.6663	.6663	.6663	.6663
		$\hat{\sigma}^2(X_{PI:S})$	=	.0011	.0007	.0006	.0002
		$\hat{\Phi}(\lambda = X_{PI:S})$	=	.872	.925	.907	.961
		$\hat{\Phi}(\lambda = .50)$	=	.920	.954	.942	.976
		$\hat{\Phi}(\lambda = .75)$	=	.889	.935	.919	.967

[a] Numbers in parentheses are G Study sample sizes; i.e., $n_p = 2432$, $n_s = 5$, $n_i = 8$.

[b] $\hat{\sigma}^2(\tau)$ is the universe score variance, which is identical to $\hat{\sigma}^2(p)$ when subtests are random, and equals $\hat{\sigma}^2(p) + \hat{\sigma}^2(pS)$ when subtests are fixed.

Figure Captions

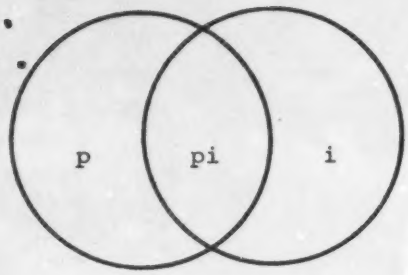
Figure 1. Venn diagram representation of mean squares and variance components for p x i design.

Figure 2. Estimates of $\Phi(\lambda)$ for various numbers of items based on Guilford (1954) data.

Figure 3. Venn diagram representation of mean squares and variance components for i:p design.

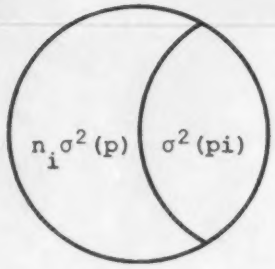
Figure 4. Venn diagram, model, and estimated random effects G study variance components for p x (i:s) design.

(A)



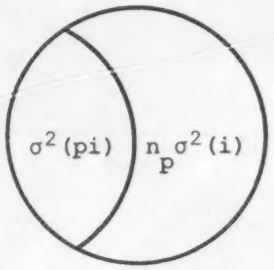
General Form of Venn Diagram for p x i Design

(B)



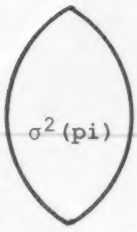
Expected Value of the Mean Square for the Effect μ_p: $EMS(p) = \sigma^2(pi) + n_i\sigma^2(p)$

(C)



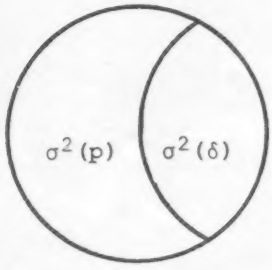
Expected Value of the Mean Square for the Effect μ_i: $EMS(i) = \sigma^2(pi) + n_p\sigma^2(i)$

(D)



Expected Value of the Mean Square for the Effect μ_{pi}: $EMS(pi) = \sigma^2(pi)$

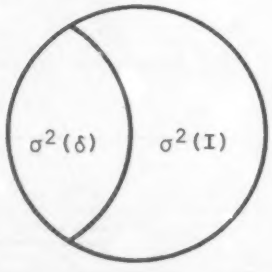
(E)



Expected Observed Score Variance:

$$\mathcal{E}\sigma^2(X) = \sigma^2(p) + \sigma^2(\delta)$$
$$= \sigma^2(p) + \sigma^2(pI)$$

(F)



$$\sigma^2(\Delta) = \sigma^2(\delta) + \sigma^2(I)$$
$$= \sigma^2(pI) + \sigma^2(I)$$

76

FIGURE 1

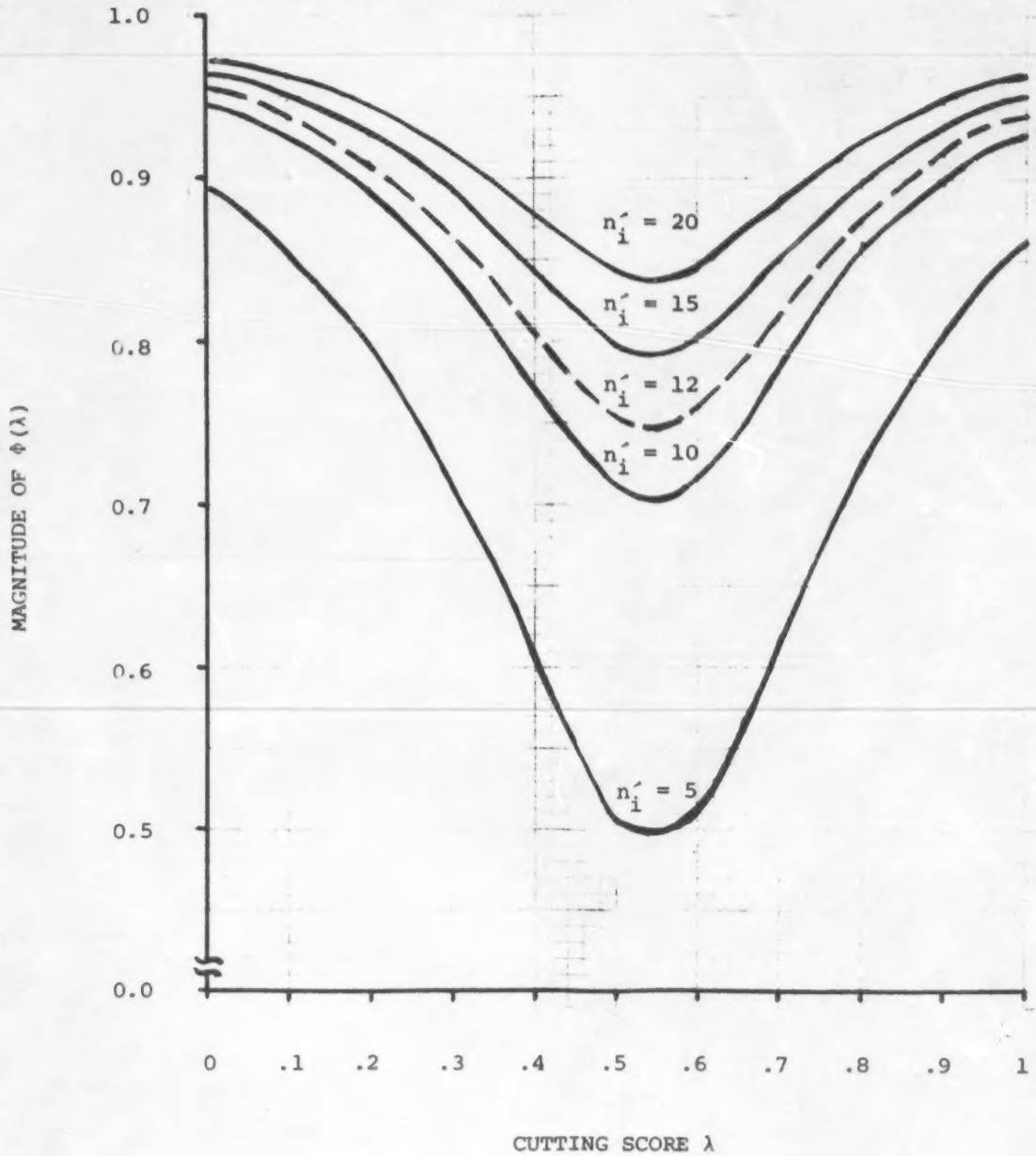
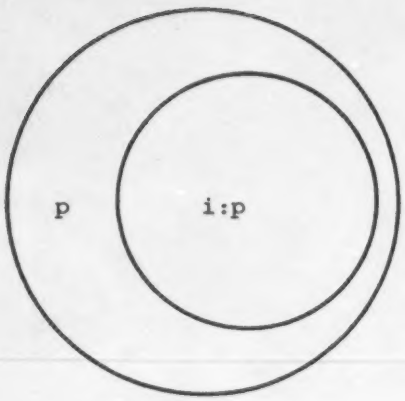

1.0
0.9
0.8
0.7
0.6
0.5
0.0
MAGNITUDE OF $\hat{\phi}(\lambda)$
$n_i' = 20$
$n_i' = 15$
$n_i' = 12$
$n_i' = 10$
$n_i' = 5$
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1
CUTTING SCORE λ

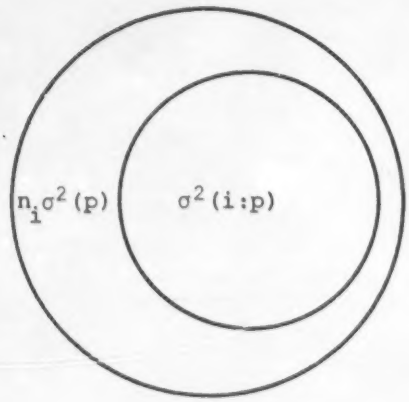
FIGURE 2

(A)



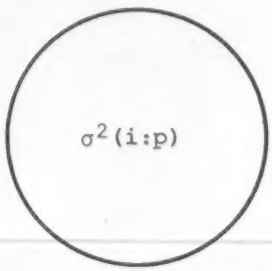
General Form of Venn Diagram for i:p Design

(B)



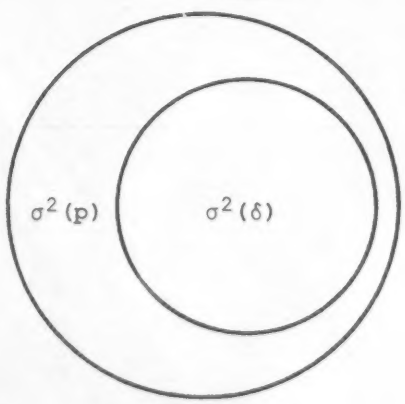
Expected Value of Mean Square for Effect $\mu_p \sim$: $EMS(p) = \sigma^2(i:p) + n_i\sigma^2(p)$

(C)



Expected Value of Mean Square for Effect $\mu_{i:p} \sim$: $EMS(i:p) = \sigma^2(i:p)$

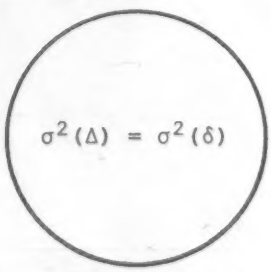
(D)



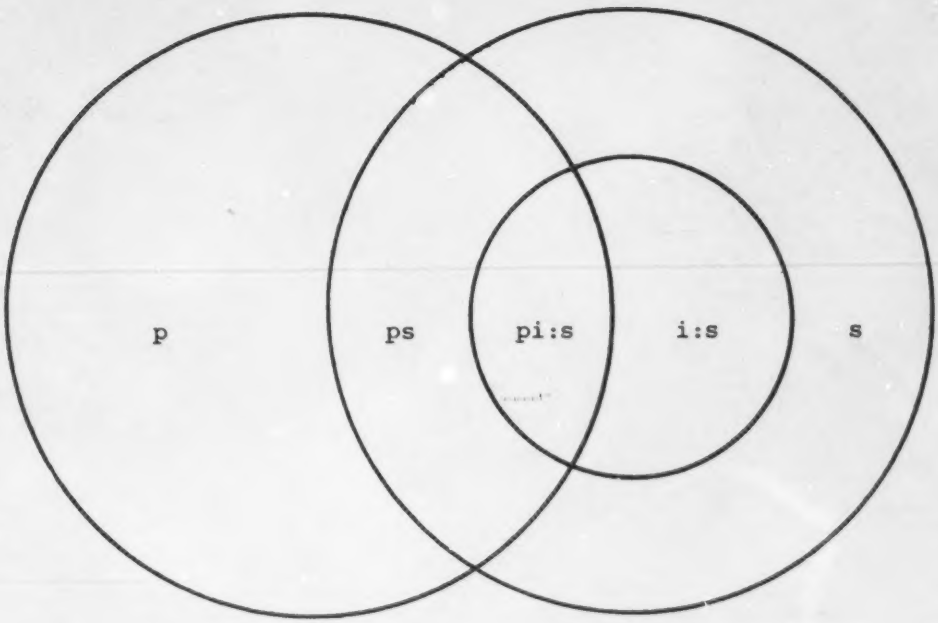
Expected Observed Score Variance:

$$\mathcal{E}\sigma^2(X) = \sigma^2(p) + \sigma^2(\delta)$$
$$= \sigma^2(p) + \sigma^2(I:p)$$

(E)



$\sigma^2(\Delta) = \sigma^2(\delta) = \sigma^2(I:p)$



Model:

$$X_{pi:s} = \mu + \mu_p^{\sim} + \mu_s^{\sim} + \mu_{i:s}^{\sim} + \mu_{ps}^{\sim} + \mu_{pi:s}^{\sim}$$

where experimental error is counfounded with $\mu_{pi:s}^{\sim}$.

Estimated G Study Variance Components:

$$\hat{\sigma}^2(p) = [MS(p) - MS(ps)]/n_i n_s$$

$$\hat{\sigma}^2(i:s) = [MS(i:s) - MS(pi:s)]/n_p$$

$$\hat{\sigma}^2(s) = [MS(s) - MS(i:s) - MS(ps) + MS(pi:s)]/n_p n_i$$

$$\hat{\sigma}^2(ps) = [MS(ps) - MS(pi:s)]/n_i$$

$$\hat{\sigma}^2(pi:s) = MS(pi:s)$$

FIGURE 4